論　文

ダンス動画への音声・視覚情報付与による低学年児童・幼児向けダンス習得支援システム

晴山　洋人[†]　　長谷川　忍[††]

A Dance Learning Support System for Lower-Grade and Preschool Children Using Audio and Visual Aids in Dance Videos

Hiroto HAREYAMA[†] and Shinobu HASEGAWA[††]

あらまし　本研究は, 見本動画を用いたダンス練習が一般化する一方, 児童・幼児が動画視聴のみで動作のタイミングや姿勢を正しく理解することが難しいという課題に着目した. 特にダンスの基本的な要素である, 姿勢を一瞬停止させる「止め」の動作に焦点を当て, 低学年児童・幼児を対象としたヒップホップダンス習得支援システムを開発し, その有効性を検証した. 本システムは, 音響・動画像解析により動画から「止め」のタイミングと姿勢を自動検出するコアエンジンと, 検出結果に基づきオノマトペ音声や視覚情報を付与する UI システムから構成される. 評価実験では, コアエンジンの最適手法を同定し, UI システムを用いて児童・幼児の練習効果を専門家が評価し, Wilcoxon の符号付順位和検定による統計的検討を行った. その結果, 短期練習では有意差は得られなかったものの, 女子のダンス経験者において「止め」の可視化が理解促進に寄与した可能性が示唆された. また, アンケートでは高い受容性が確認され, 特に視覚情報の有効性が顕著であった. 以上より, 本研究は従来研究で注目されなかった「止め」の自動検出技術を応用し, 児童・幼児向けダンス支援の新たな可能性を示すものである.

キーワード　ダンス練習, 自動検出, 児童・幼児, 音声付与, 視覚情報付与

1. はじめに

1.1 研究の背景

近年, ダンスは教育・スポーツ・メディアなど多様な分野で注目を集めている. 小学校ではリズムダンスが「表現運動」として取り入れられており [1], 2012 年には中学校の保健体育でダンスが必修化された [2]. また, 2024 年のパリ五輪ではブレイキン（ブレイクダンス）が正式種目として採用されている [3]. さらに, 2020 年には日本初のプロダンスリーグ「D.LEAGUE」[4] が開幕し, ダンスがプロスポーツとしての道を歩み始めた. これらの動向は, ダンスが社会的に広く認知され, さまざまな世代に浸透していることを示している. このような社会的関心の高まりを背景に, 児童・幼児におけるヒップホップダンスの人口が増加している. 笹川スポーツ財団の調査 [5] によれば, 4〜11 歳 (標本数:2,400 人) を対象とした週 1 回以上のヒップホップダンス実施率は増加傾向にあり, 図 1 1 に示すように 2013 年から 2023 年にかけて上昇し, 2.7% から 4.1% に増加した. 住民基本台帳人口 (4 11 歳)(人) に上記実施率を乗じて算出した推計人口は 24 万人から 33 万人に増加している.

ヒップホップダンスの基本的なリズムの取り方は, 膝を屈伸させて沈み込む「ダウン」と膝を伸ばして体を引き上げる「アップ」である [6]. また, 基本的な動きの一つである「上肢のウェーブ動作」はウェーブを伝搬させる速度が一定, 振幅が一定で正弦波のように滑らかである. つまり, リズムをとりながら周期的に身体を「止め」る動作を循環的に行うのがヒップホップダンスの特徴である.

児童・幼児はヒップホップダンス（以下, ダンス）を基本的な動きから段階的に習得していく. 基本的な動きは, 姿勢を一瞬静止させる「止め」の連続で構成され

[†]*
*
[††]北陸先端科学技術大学院大学　先端科学技術研究科
Japan Advanced Institute of Science and Technology
DOI:10.14923/transfunj.???????????

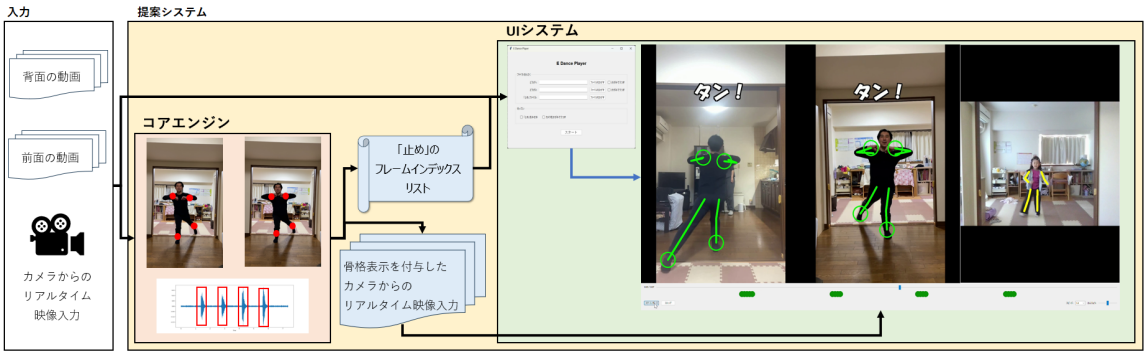

図 1　提案手法のシステム概要

見本となる (ダンス指導者の) 動画からコアエンジンを通して「止め」のフレームインデックスリストを抽出する. また, 同じ動画と同期して撮影したその背面の動画を UI システムで指定して入力する. 最後に, 見本の前面・背面の動画と, カメラからのリアルタイム入力にコアエンジンの処理を合わせた動画を UI システムで表示する.

ており, 各ステップのカウントにおける「止め」の動きを通じて振付を学んでいく. この「止め」の動きは, 上級者においてもキレやタメといった高度な表現を生み出す重要な技法である. [7] [6] たとえば, 次の動きへの移行前にポーズをとることで安定感や表現を強調する「Pose」や, 音楽のビートに合わせて身体を一瞬固定させる「Hit」といった手法がある.

こうした動きを児童・幼児がダンスを習得する際には, ダンス教室などで指導者の動きを模倣することが一般的である. 内山はダンス学習の際に最もオーソドックスな方法は「模倣」であると述べている [8]. また飯野らは上級者の指導の下, 鏡で自分の姿を見ながら修正するか, DVD などの映像を見て真似るかのいずれかが主な練習方法であると主張している [9]. ダンス教室では, 指導者は児童・幼児の前に立ち, 鏡越しに後ろ向きで踊る. その結果, 児童・幼児は指導者の背面の動きと鏡に映った左右反転した動きを同時に観察しながら, 身体を動かし, 振付を学んでいくことが行われている. またスマートフォンが普及した現代では, 指導者の動きを撮影し, その動画を家庭で確認しながら自主練習することも推奨されている. ダンスの技術を向上させるためには, 日々の練習が極めて重要である. 中でも, 動画を活用した自主練習は技術向上だけでなく, ダンスを楽しむうえでも欠かせない活動であると考えられる.

しかし, 動画を視聴するだけでは, 児童・幼児にとって「止め」のタイミングやその姿勢を正しく理解することは難しいと考えられる. 実際のダンス指導の現場では, 「タン・タン」などのオノマトペ（擬音語）を用いて動きのタイミングを伝え「止め」の姿勢について指導者が説明を交えて指導し, それを児童・幼児が実践することで, 段階的に振付を習得しているのが現状である.

1.2　研究の目的

そこで本研究では, ヒップホップダンスを対象とし, ヒップホップダンスにおける「止め」の動きに着目する. 動画上の「止め」の瞬間に音声・視覚情報を付与することで, 児童・幼児のダンス習得を支援することを目的とする. まず, 児童・幼児が見本動画から「止め」の動きを自動で検出できるシステムを開発する. 次に, 「止め」の動きに対して, 音声, 記号, 文字といった情報を付与することで, 児童・幼児にとってより分かりやすく振付を理解できるようにするシステムを構築する.

さらに, 上記のシステムを用いて, 音声・視覚情報を付与した動画と, 付与していない動画を用いた場合で, 児童・幼児のダンス習得に差が生じるかを検証する比較実験を行う.

1.3　本論文の構成

本論文の構成は以下の通りである. まず第 2 章では本研究に関連する先行研究について述べる. 先行研究は大きく教育的観点, 支援システム関連, 骨格検出関連に大別される. 第 3 章では本研究の提案手法について述べる. また, 本研究の要件について述べ, 課題を明確化する. 第 4 章では提案手法の実装について述べる. 主に「止め」の動きを検出するシステムであるコアエンジンと児童・幼児に「止め」の動きの理解を促す UI シ

ステムについて述べる. 第 5 章では提案システムを使用した実験及び評価について述べる. 第 6 章では最終的な成果と今後の課題について述べる.

2. 関 連 研 究

本章では, ダンス習得に関する先行研究について, 教育的観点からの研究とシステムを活用した研究, 及び「止め」の検出の際に使用する骨格検出に関する研究に大別して概説する.

2.1 教育的観点からのダンス習得に関する研究

教育的観点からのダンス習得に関する研究では, 幼児や児童を対象とした効果的な指導法や学習過程に関する実践的な知見が報告されている. たとえば, 指導段階を明確に区分して考察する研究 [10] や, 指導者と児童が「よい動き」について共通理解を持つことで技能向上を図る研究 [11], 特定のダンス技法の習得方法に焦点を当てた研究 [12] などがある.

亀山は幼稚園年長児（5〜6 歳）を対象にリズムダンスの習得過程を「①導入期, ②練習期, ③葛藤期, ④変化期, ⑤発表期」の 5 段階に分類している [10]. 特に③葛藤期において, 子どもが音やリズムを感じながら, 自身の身体を通して動きを獲得していく過程が重要であることが示されている.

湯浅らは, 小学校低学年におけるリズム遊びの指導において, 「よい動き」に対する合意形成を行うことで, 児童の動作パフォーマンスの向上が認められたと報告している [11]. さらに湯浅は, 小学校 2 年生（7〜8 歳）に対するリズム系ダンス授業の実践を通じて, 児童同士が「よい動き」を共通理解することで, 着目する身体部位が明確化され, さらに他者と関わりながら踊ることで, ダンス技能の習得が促進されると主張している [12].

本研究に関連の深い研究として, 高田は, 幼児期および児童期初期の発達段階に応じたダンス実践を報告しており, 幼児におけるダンス習得には「動作の分割（上半身と下半身を別々に動かすこと）」や「音楽に合わせて踊る時間の確保」が重要であるとしている [13]. ただし, 各ステップにおける「止め」の動きに関する分割については具体的に言及されていない.

天野らは, 「言語のみ」「オノマトペのみ」「言語とオノマトペの併用」「カウントのみ」の 4 種類の指導方法が動作習得過程に与える影響について, 比較実験を通じて検討した. これらの指導方法はいずれも, 動画に音声を付加する形で提示された. その結果, 初めから「言語とオノマトペ」を併用して指導する場合, 学習者が情報過多になる可能性が示唆された. また, 初見の振り付けを学習する際には, まず全体の流れを把握し, その後に詳細な動作を段階的に習得するという学習プロセスが有効であることが示唆された [14]. また, [14] の中で斎藤ら [15] の研究との違いが論じられている. 斎藤ら [15] の研究は動画に「音声のみ」, 「文字のみ」で動画にオノマトペを付与するもので, [14] ではその「音声のみ」のオノマトペ付与をより深く研究したものと言える. [14] の研究の中で, 「音声のみ」のオノマトペ付与効果が限定的であったことから, 斎藤ら [15] の「文字のみ」オノマトペ付与でダンス習得効果が向上したことと比較し, 動画による学習においては, 聴覚情報よりも視覚情報の方が学習に強く影響を及ぼす可能性を主張している. これは, 言語やオノマトペを文字として画面表示することがダンス習得に影響を及ぼすことを示唆している.

最後に, 本研究と関連の深い先行研究との共通点及び相違点を以下に示す.

＜共通点＞

- 児童・幼児のダンス習得において動作の分割が重要であることが報告されている.
- ダンス習得において, オノマトペを音声として付与する影響を報告している.

＜相違点＞

- 動画の分割について述べられているが, 「止め」の動作に注目した分割については言及されていない.
- [14] ではオノマトペの音声付与のみの影響について言及しており, [15] ではオノマトペの視覚提示の可能性が述べられている. しかし, 音声と文字表示の両方を付与した場合についての言及はない.

2.2 システムによるダンス習得に関する研究

システムを活用したダンス習得支援に関する研究では, さまざまな提案がなされている. これらには, ダンス動画への視覚情報の付加による学習効果の検証 [15] や, 動画の自動分割技術の応用 [16] などが含まれる.

田中らは, ストリートダンスの動作データを解析し, 動きの特徴を抽出した上で上級者との比較を行い, その結果を直感的に把握できる形で可視化するシステムを提案している [17].

山内らは, Kinect とワイヤレスマウスを組み合わせたダンス学習支援システムを提案し, マウスによるタッピング操作を用いてリズム感の習得度合を定量的に判定可能とした [18].

西脇らは, 学習意欲が高くないユーザーでもダンスの基礎やステップを習得できるよう, 習熟度に応じて動作判定やフィードバックを変化させるシステムを開発した [19]. また, ユーザーが踊りながら間違いを修正できるよう, リアルタイムで音声によるフィードバックを行った.

何らは, 適応型ダンス練習支援システム「FreeDance」を提案し, 三面壁型透明スクリーンを用いることで高い没入感を実現し, 学習者の動機づけを促進している [20].

戸山らは, 撮影したダンス動画を加速度データに基づいて動作の単位に自動で分割し, その単位ごとに繰り返し再生可能なインタラクティブチュートリアルを実現している [21]. 実験の結果, 初心者にとっては未加工の動画よりも本システムによる学習の方が振り付けを覚えやすいことが示された.

本研究と関連の深い先行研究として, 斎藤らは, 漫画風オノマトペをダンス動画に視覚的に付与することで, ダンス習得効果が向上することを示した [15]. 本研究においても同様の視覚的オノマトペの付加を行っているが, 音声と文字の両方を付与した際の比較については言及がない.

また, Endo らは, ダンス動画から振り付けの短時間の動きを自動で分割する手法を提案しており, キーポイント間の速度変化を視覚特徴量として利用している [16]. 本研究のように「止め」の動きの検出については言及されていない点, および幼児・児童向けの UI 設計には対応していない点で差異がある.

さらに, Anderson らは, 自己学習を支援する AR ミラー型インターフェース「YouMove」を提案している [22].「YouMove」はユーザーの姿勢をリアルタイムに解析して見本との違いを可視化し, 動作の一時停止や繰り返し再生などの機能を備えている. 本研究でも同様に骨格推定や動画の一時停止・繰り返し再生機能を実装しているが, 見本動作の表示方法としては, 重ね合わせ表示ではなく並行表示を採用している点で異なっている.

最後に, 本研究と関連の深い先行研究との共通点及び相違点を以下に示す.

＜共通点＞

- 動画にオノマトペを付与した場合の効果を示している.
- ダンスの振り付けを自動で分割する手法について述べている.
- 自己学習を支援するインターフェースを提案している.

＜相違点＞

- 動画にオノマトペの音声と文字両方を付与した場合の効果については述べられていない.
- 「止め」の分割についての言及はない.
- 見本動作の表示方法として, 重ね合わせではなく, 並行表示を採用している.

2.3 骨格検出に関する研究

骨格検出（Pose Estimation）は, 画像または映像内の人体の関節位置を特定する技術である. 人体の動きや姿勢を機械が理解するための基盤技術として, 多くの研究が行われてきた. 初期には特徴点を用いた機械学習による分類が行われたが, 特にディープラーニングの進展により, リアルタイムで高精度な骨格検出が可能となり, さまざまな実用化が進んでいる.

Cao らによる OpenPose [23] は, 初めてマルチパーソン骨格検出をリアルタイムで実現したオープンソースのシステムである. Part Affinity Fields（PAFs）と呼ばれる空間的なベクトル場を導入することで, 複数人物に対する 2 次元姿勢推定を高精度でリアルタイムに行うことができる. ただし, リアルタイム処理については GPU を使用した場合に限り実現可能であることが報告されており, CPU のみでのリアルタイム実行は困難である.

Bazarevsky らによる MediaPipe [24] は, Google が開発したオープンソースのマルチモーダル機械学習フレームワークである. リアルタイムな画像処理および機械学習パイプラインの構築を支援するプラットフォームで, 手や顔の検出, 姿勢推定などの高精度な機能を備えている. 組み込み機器やスマートフォンでのリアルタイム推論を目的とした軽量なライブラリであり, CPU でもリアルタイムに実行可能である. 少ない計算リソースでも一定の精度で実行できることが確認できたため, 本研究では MediaPipe を利用することとする.

Xu らによる ViTPose [25] は, 人体姿勢推定のための新しいアーキテクチャであり, Vision Transformer（ViT [26]）を活用した手法である. このモデルは, 従来の畳み込みニューラルネットワーク（CNN）ベースのアプローチに比べ, 視覚的特徴の処理において高い効率性と精度を発揮する. ViTPose は, 画像内の各関節位置を予測するために, トランスフォーマーの自己注意メカニズムを活用し, 物体の局所的およびグローバルな特

徴を効果的に学習する. しかしながら, 環境構築が複雑で, 要求リソースも大きいことから本実装では採用しなかった.

3. 提案手法

3.1 アーキテクチャ

本研究は, 児童・幼児のダンス習得支援を目的として, 各ステップにおける「止め」の動作に着目し, 視覚および音声情報を付加することにより,「止め」の姿勢およびタイミングの理解を促進するシステムの提案を行うものである. 特に,「止め」の動作を視覚的に明確化し, 音声的な手がかりと組み合わせることで, 児童・幼児のダンス学習を効果的に支援することを目指す.

図 1 に提案システムのアーキテクチャ図を示す. 入力として見本となる (ダンス指導者の) 動画から「止め」のフレームインデックスリストを抽出する. また, 同じ動画と同期して撮影したその背面の動画を UI システムで指定して入力する. 入力動画の「止め」のフレームに音声・視覚情報を付与し, カメラからのリアルタイム入力を合わせて UI システムで表示する. カメラからの表示には「止め」の抽出と同じアルゴリズムを用いて視覚情報を付与する.

提案システムは, 以下の 2 つの主要な機能で構成される.

- コアエンジン：見本となる (ダンス指導者の) ダンス動画を入力とし, 音響情報および動画像情報を解析することで,「止め」の動作を自動的に検出する.
- ユーザーインターフェースシステム (UI システム)：検出された「止め」のフレームに対して, 視覚および音声情報を付与することで, 児童・幼児が動作の内容とタイミングを理解しやすくなるよう支援する.

コアエンジンでは, 入力されたダンス動画から音響情報および動画像情報を抽出する. 音響情報に関しては, 周期的な音のピークを検出し,「止め」のタイミングの候補フレームを抽出する. 一方, 動画像情報においては, 骨格推定によりダンス上級者の手首および足首のキーポイントを検出し, 各フレーム間における移動速度がゼロとなる箇所を「止め」の姿勢候補として抽出する. これら双方の候補が一致する動画フレームを「止め」の動作として確定する.

UI システムでは, 検出された「止め」の動画フレームに対して, オノマトペによる音声情報および記号・文字による視覚情報を重ね合わせる. また, カメラからのリアルタイム入力にコアエンジンと同じアルゴリズムを用いて骨格情報を視覚的に付与する. これにより, 児童・幼児は視覚と聴覚の両面から「止め」のタイミングと姿勢を直感的に理解することが可能となる.

3.2 要　　件

本研究で解決すべき課題は以下の通りである.

3.2.1 コアエンジンの課題

- 音響情報により「止め」のタイミングを検出する.
- 動画像情報により「止め」のタイミングの姿勢を検出する.

3.2.2 UI システムの課題

- 検出した「止め」の動作に音声・視覚情報を付与する.

4. 実　　装

4.1 コアエンジン

コアエンジンの入力と出力は以下である.

- 入力：動画 (.mp4) ファイル
- 出力：「止め」の動作を行っている動画フレームインデックスのリスト

コアエンジンでは, 音響情報で検出したフレームと動画像情報で検出したフレームの共通フレームを「止め」の動作を行っている動画フレームとして検出し, その動画フレームのインデックスのリストを出力する. また, ダンス経験者とのディスカッションの中で,「止ま」り始める部分についても「止め」の動作とするとよいとのアドバイスを受け, 共通フレームが連続 3 フレーム以下の場合は, 1 フレーム前のフレームも「止め」のフレームとした.

本研究での入力動画ファイルの条件は以下の通りとする.

- 入力動画ファイルは BPM(Beats Per Minute)=90 のダンス振り付けを撮影した動画である.
- 入力動画ファイルの fps は 30.0 である.
- 入力動画ファイルはメトロノーム音が鳴っている中で撮影した動画である.
- メトロノーム音が鳴っているタイミングが「止め」のタイミングの候補となる.
- メトロノーム音が動画内の音響情報において主要な要素を占めており, 他には足音などの微小な環

境音がわずかに含まれるのみである.

4.1.1 音響情報による「止め」のタイミングの検出

音響情報による「止め」のタイミング検出手法として以下の 2 手法を実装する.

＜振幅特徴ベース＞

入力動画の周期的な音響情報 (メトロノーム音) にて音の振幅がピーク (局所最大値) となる動画のフレーム番号を検出する. 音の振幅は動画内のフレーム数を N, 動画 1 フレーム単位での音の振幅 $x_i (i = 1, \ldots, N)$ とした時, 以下 2 つの条件を満たすフレームをピークとして検出する.

閾値 (振幅の高さ h の条件)：$h = \mu + \sigma$

ここで μ と σ は以下の式で表される.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \tag{2}$$

閾値で検出したフレームの前後 1 フレームもピークとする.

＜自己相関関数ベース＞

自己相関関数を用いた手法はまず対象とする音声ファイルから単一チャネル (モノラル) として信号 $y[n]$ を抽出する. サンプリング周波数を f_s とする. 信号の周期的構造を明らかにするために, 自己相関関数 $R[k]$ を用いる. これは以下の式で定義される.

$$R[k] = \sum_{n=0}^{N-1-k} y[n] \cdot y[n+k] \quad (0 \le k \le N) \tag{3}$$

ここで N は信号の長さである. 次に自己相関関数 $R[k]$ から peak 位置を検出する. このとき, メトロノームが打つ間隔 (BPM=90 より 0.5 秒程度) に基づき, peak 間の最小処理を制限することで可検出を防ぐ.

$$P_{\text{auto}} = FindPeaks(R[k], distance = f_s \cdot 0.5) \tag{4}$$

周期推定値は, 検出されたピーク間の中央値を用いて次のように求める.

$$T_{\text{est}} = median(\Delta P_{\text{auto}}) \tag{5}$$

元の音響信号 $y[n]$ 上で音の peak 位置を求める. peak 検出においては, 次の 2 つの条件を課す:

- 信号の振幅が最大値の 50% 以上であること (=メトロノーム以外の微小ノイズを除外)
- 推定周期の 80% 以上の間隔で peak を制限 (=1 打で複数の peak を検出しない)

$$P_{\text{auto}} = FindPeaks(y[n], \; height \ge 0.5 \cdot max(y), \; distance \ge 0.8 \cdot T_{\text{est}}) \tag{6}$$

ここで, P_{auto} は音声サンプルインデックスのリストである.

動画のフレームレートを f_{fps}[fps] としたとき, サンプル番号 n は以下の式により対応するフレーム番号 f に変換される.

$$f = \lfloor \frac{n}{f_s} \cdot f_{\text{fps}} \rfloor \tag{7}$$

以上により, 自己相関関数を用いてメトロノームが鳴るフレーム番号のリストを得る. peak 検出には Scipy [27] の find_peaks 関数を用いる.

4.1.2 動画像情報による「止め」のタイミングの姿勢検出

動画像情報による「止め」のタイミング検出手法として以下の 3 手法を実装する. これに先立ち, 前処理として各キーポイントの速度情報を算出する。まず, 骨格検出モデル MediaPipe [24] を用いて左右手首・足首のキーポイントを検出する. 次に, 検出したキーポイントの動画フレーム間速度を算出する. フレーム t で検出した i 番目のキーポイントの位置 (x_i, y_i) を $k_i(t) \in \mathbb{R}^2$, 速度 $v(t) \in \mathbb{R}^{(4\times2)}$ の i 番目の要素 $v_i(t) \in \mathbb{R}^2$ を以下の式で求める.

$$v_i(t) = |k_i(t) - k_i(t-1)| \tag{8}$$

＜閾値アルゴリズム＞

上式で算出した速度 $v(t)$ が閾値以下のフレームを「止め」のタイミングの姿勢とした. 本研究では後述する予備実験により, 閾値の値を 10[pixel/frame] とする.

＜ *Peak* 検出アルゴリズム＞

peak 検出では, まず各キーポイントの速度系列データ $(v_i(1), v_i(2), \ldots, v_i(n))$ に対して, 最小値 v_{imin} 及び最大値 v_{imax} を用いた min-max 正規化を行う.

$$\hat{v}_i(t) = \begin{cases} \dfrac{v_i(t) - v_{i\,\min}}{v_{i\,\max} - v_{i\,\min}} & \text{if } v_{i\,\max} > v_{i\,\min}, \\ & t = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

また, $\hat{v}_i(t)$ の t に対して平均をとることで統合信号 $s_t = (s_1, s_2, \ldots, s_n)$ を構成する：

$$s_t = \frac{1}{d}\sum_{j=1}^{d} \hat{v}_i(t), \text{for } t = 1, 2, \ldots, n \tag{10}$$

ここで d は左右手首・足首のキーポイント位置 (x, y) であるため, $d = 8$ となる. 統合信号を再び min-max 正規化した後, peak 検出アルゴリズムを適用する. 本実装では peak 検出アルゴリズムとして Scipy [27] の find_peaks 関数を用いた. パラメータは peak 間の最小距離を 15 フレーム (fps30 で約 0.5 秒), peak の顕著性 (突出度) をノイズ除去のため 0.3 とした.

$$\begin{aligned}\mathcal{P} = \text{FindPeaks}(s_t, \\ distance = 15, \\ prominence = 0.3)\end{aligned} \tag{11}$$

得られた peak 位置の集合を $\mathcal{P} = (p_1, p_2, \ldots, p_k)$ とすると, 各 peak $p \in \mathcal{P}$ の周辺 $[p - w, p + w]$ を動いている区間とみなし, ブール配列 $h_t = (h_1, h_2, \ldots, h_n)$ を次のように定義する.

$$h_t = \begin{cases} 1 & \text{if } \exists p \in \mathcal{P}, |t - p| \le w \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

ここで $w = \lfloor \frac{distance}{2} \rfloor = 7$ である. 最終的に $h_i = 0$ すなわち動いている区間に属さないインデックス t を「止め」の区間とみなし, それらのインデックス集合を $\mathcal{L}$ として抽出する.

$$\mathcal{L} = \{t \in \{1, 2, \ldots, n\} \mid h_t = 0\} \tag{13}$$

＜ *k-means* アルゴリズム＞

k-means のクラスタリングでは, まず各キーポイントの速度系列データ (速度 $v_i(1), v_i(2) \ldots, v_i(n)$) に対して, 以下のように標準化を行い, 平均 0, 分散 1 の正規化済みデータを得る:

$$\tilde{v}_i(t) = \frac{v_i(t) - \mu_i}{\sigma_i} (1 \le t \le n, 1 \le i \le d) \tag{14}$$

ここで $d = 8$ である. また, μ_i 及び σ_i は各キーポイントの要素 (左右手首・足首の (x, y)) の平均と分散である. 標準化された $\tilde{v}(t)$ に対して, クラスタ数 $k = 2$ の k-means クラスタリングを実行し, 各サンプル t が属するクラスタラベル $c_i \in \{0, 1\}$ を得る.

$$c_i = \text{argmin}_{k \in \{0,1\}} \| \tilde{v}(t) - \mu_k \|^2 \tag{15}$$

ただし, μ_k はクラスタ k の中心である. クラスタ中心の各次元の平均値を計算し, 平均が大きいクラスタを「動作クラスタ」, 小さい方を「「止め」クラスタ」とする. クラスタ k の平均値は以下で定義される.:

$$m_k = \frac{1}{d}\sum_{i=1}^{d} \mu_{k,j} \tag{16}$$

ここで, $\mu_{k,j}$ はクラスタ k の i 番目の次元の中心値である. 平均 m_0 と平均 m_1 を比較し, $m_1 > m_0$ の場合はクラスタ 1 を「動作クラスタ」, そうでなければクラスタ 0 を「動作クラスタ」とする. クラスタラベルが「動作クラスタ」ではないサンプル t を抽出し, 昇順でソートしたものが「止め」のフレームインデックスリストである.

4.2 UI システム

UI システムの入力と出力は以下である. また, 図 4 1, 図 2 の通り実装した.

- 入力：
  - 同期させて撮影した見本のダンス動画ファイル (.mp4)
    - 前面から撮影された動画ファイル (図 4 1(a))
    - 背面から撮影された動画ファイル (図 4 1 (a))
  - コアエンジンで検出した「止め」の動作を行っている動画フレームインデックスのリスト (図 4 1 (c))
  - カメラからの動画像 (リアルタイム表示)(図 2(c))
- 出力：
  - 入力された動画ファイルに音声・視覚情報を付与した動画 (図 2 (a), (b))
  - カメラからの動画像に視覚情報を付与したリアルタイム表示 (図 2 (j))

上記の”音声情報の付与”とは以下を行うことである.

- 「止め」のタイミングでオノマトペの音声情報を付与する.

上記の”視覚情報の付与”とは以下を行うことである.

- ☒ 左右の肩から手首, 腰から足首までに骨格表示を行う. 骨格表示は「止め」の姿勢では緑色にな

り, それ以外は黄色になる. (図 2 (h))
- 「止め」の姿勢の際に左右手首・足首に丸の図形付与 (図 2(h))
- 「止め」の姿勢の際に漫画風オノマトペの文字を付与 (図 2 (i))
- シークバーの「止め」のタイミングに緑の印を付与 (図 2 (g))

また, UI システムは以下の機能を持つ.
- 動画の再生・停止機能 (図 2(d))
- 再生速度の変更機能 (変更粒度は動画オリジナルの速度を 1.0 として 0.25 刻みに 0.25 2.0 まで) (図 2(e))
- 音量調整機能 (図 2(f))
- 入力動画及びカメラ表示の左右反転機能 (図 4 1(b, e))
- 音声・視覚情報の付与有無を切り替える機能 (図 4 1(d))
- 動画を開始する際, 5 秒待つ機能

4.3 環　　境

本実装では以下の環境及びプログラミング言語にて実装を行った.　また Python のライブラリについて主要なものを以下に列挙する.

OS:	Windows 11 Home 24H2
CPU:	11th Gen Intel(R) Core(TM) i7-1195G7 @ 2.90GHz
RAM:	16.0GB
language:	Python(v3.11.9 embed-amd64)
library:	MediaPipe 0.10.18, NumPy 1.26.4, Pandas 2.2.3, SciPy 1.14.1, scikit-learn 1.6.0, python-vlc 3.0.21203, opencv-contrib-python 4.10.0.84, Pillow 10.4.0, Tkinter 8.6.12, ffmpeg-python 0.2.0, librosa 0.10.2.post1

5. 実験・評価

5.1 コアエンジンの実験・評価

コアエンジンで適切に「止め」を検出できるかを確認することを目的に, 以下の実験・評価を行った.

5.1.1 コアエンジンの実験

1. 評価用の動画を用意する. 本実験では 5 つの動画を評価用に用いた.
2. ダンス経験者監修のもと「止め」のフレームにアノテーションを行った. 具体的には各動画フレームを一枚ずつ画像に分割し, 「止め」のフレームだと考える動画フレームインデックスのリストを作成した.
3. コアエンジンで「止め」のフレームを推定した.
4. 推定した「止め」のフレームとアノテーションした「止め」のフレームが合致するかダイスインデックス (Sørensen-Dice coefficient) によって評価した.

図 2　UI システム動作画面

(a) 背面動画, (b) 前面動画, (c) カメラ表示, (d) 動画の再生・停止, (e) 動画再生速度変更, (f) 音量調整, (g)「止め」タイミング印, (h)「止め」の際の骨格・図形表示, (i) オノマトペ表示, (j) カメラ動画像へのリアルタイム骨格表示

ここで, 推定した「止め」のフレームを A, アノテーションした「止め」のフレームを B とするとき, ダイスインデックスの式は以下である.

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{17}$$

5.1.2 コアエンジンの評価

上記ダイスインデックスを用いた音響情報による「止め」の検出手法の比較と, 動画像情報による「止め」の検出手法の比較評価をそれぞれ行った. また, 動画像情報による「止め」の検出手法の評価では, 閾値アルゴリズムの閾値評価, 共通フレームの前後にフレームを拡張してダイスインデックスの値に変化があるか評価を行った. さらに, ダイスインデックスが高い事例と低い事例に関する考察も行った.

＜表の文言整理＞

ここで, 表の文言は以下の通りである.
- method : 使用した手法. threshold が提案手法での評価結果.
- win_num : 共通フレームの前後に拡張したフレー

表 1　コアエンジン評価 音響情報による「止め」の検出手法比較

	Sample1		Sample2		Sample3		Sample4		Sample5	
method	threshold		threshold		threshold		threshold		threshold	
audio	-	corr	-	corr	-	corr	-	corr	-	corr
win_num	win1	win1	win1	win1	win1	win1	win1	win1	win1	win1
GT	12	12	21	21	18	18	26	26	21	21
cnt	34	23	32	20	26	23	43	28	28	19
TP	9	7	8	2	10	6	6	4	14	8
FP	25	16	24	18	16	17	37	24	14	11
FN	3	5	13	19	8	12	20	22	7	13
TN	175	184	167	173	178	177	149	162	177	180
Acc	0.868	0.901	0.825	0.825	0.887	0.863	0.731	0.783	0.901	0.887
Recall	0.750	0.583	0.381	0.095	0.556	0.333	0.231	0.154	0.667	0.381
Precision	0.265	0.304	0.250	0.100	0.385	0.261	0.140	0.143	0.500	0.421
Dice	0.391	0.400	0.302	0.098	0.455	0.293	0.174	0.148	0.571	0.400

ムの数.

- GT : Ground truth. アノテーションしたフレームの数.
- cnt : 検出したフレーム数.
- TP : True Positive. 検出したフレームと GT が合致した数.
- FP : False Positive. 検出したフレームと GT が合致しなかった数.
- FN : False Negative. 検出しなかったフレームが GT であった数
- TN : True Negative. 検出しなかったフレームが GT でなかった数.
- Acc : Accuracy. 精度の式は以下である.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

- Recall : 再現率の式は以下である.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

- Precision : 適合率の式は以下である.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{20}$$

- Dice : ダイスインデックス.

＜音響情報による「止め」の検出手法の評価結果＞

表 1 に示された結果から, 振幅特徴ベース手法はダイスインデックスにおいて, 動画 5 つの内 4 つで自己相関関数ベース手法よりも高い値を示した. 特に動画 2 については提案手法 0.302, 自己相関関数を用いた手法が 0.098 と大きな差が確認できる. しかし, 動画 1 では自己相関関数を用いた手法の方がダイスインデックスの値が高く, 動画 4 においては 0.26 ポイントしか差がない. 相対的にコアエンジンの評価実験において, 振幅特徴ベース手法の方が高いダイスインデックスを示すことから, UI システムの実験では振幅特徴ベース手法を用いることとした. ただ, これはあくまでも今回の設定での評価であり, 動画サンプルの数を多くした場合や, 動画サンプル内の環境音やノイズが多い場合を考慮すると, よりロバストな環境でどちらを選ぶべきなのかは今後の課題である.

また, 動画像情報による「止め」の検出手法の評価にて後述するが, 拡張フレーム数を win1 に, 検出アルゴリズムは閾値アルゴリズムにて評価を行った.

＜動画像情報による「止め」の検出手法 *(*閾値アルゴリズム*)* の閾値結果＞

予備実験として, 閾値アルゴリズムの閾値をいくつか試し, 閾値を決定した. 表 5 2 表 5 6 に示した通り, ダイスインデックスの値が最も高い 10 を閾値とした. この時, 拡張フレーム数は win1 とした.

＜動画像情報による「止め」の検出手法の評価結果＞

表 5 7～表 5 11 に示された結果から, 閾値アルゴリズムは拡張フレーム数 win1 win3 のいずれにおいてもダイスインデックスが 0.174 0.571 となっており, peak 検出アルゴリズムの 0.041 0.409, k-means アルゴリズムの 0.105 0.484 に比べて相対的に高い値を示している. 特にダイスインデックスが最も高い値を示す手法は動画 1 5 において全て閾値アルゴリズムとなっている. また, Accuracy, Recall, Precision についても他の手法と比べ閾値アルゴリズムは相対的に高い値を示している. ただ, Precision については動画 1 5 において閾値アルゴリズムでも 0.140 0.500 となっており誤検出の

割合が高いことが課題として挙げられる.

提案手法での拡張フレーム数を変化させて行った実験では, 動画 1 5 に対して win1 が最もダイスインデックスが高い場合が多かった. そのため, UI システムの実験では win1 の設定にて実験を行う.

＜コアエンジン評価の総括＞

音響情報及び動画像情報による検出では, 各動画でダイスインデックスの値が 0.174 0.571 となり, また GT と TP から「止め」のアノテーションの約 1/2 を検出できたことが示され, コアエンジンの課題に寄与できたと考える. しかし, 以下考察の通り, 「止め」を検出しにくい動画への対応が今後の課題である.

＜ダイスインデックスが高い事例と低い事例に関する考察＞

今回の実験では, 動画 4 がどの手法も総じてダイスインデックスの値が低く, 動画 5 のダイスインデックスの値が最も高かった. 動画 4 のダンスは「止め」の動画の繰り返しでありつつも, やや流れるような動きであったため, 「止め」を検出しにくかったと考えられる. また, 動画 4 のダンスは身体を大きく使う振付になっており, 反動をつけるために次の動作への予備動作が比較的大きくなったと考えられる. 対して, 動画 5 では 1 拍 1 拍を「止め」, 身体を使う範囲が比較的狭く予備動作も少ないことから, 検出がしやすかったと考えられる. 同様の考察は [16] でもなされており, 拍の動きにアクセントが来る振りの場合は [16] の論文で議論されている動画分割が行いやすく, 反対に柔らかく流れるような振付に対しては分割が難しかったと述べられている.

5.2 UI システムの実験・評価

UI システムを用いて児童・幼児にダンス練習を実施してもらい, 音声・視覚情報の有無でダンスの動きとリズムの習得度に変化があるか評価した. また, 音声・視覚情報の付与がダンス習得に有効に働いた群はどのような背景属性を持つか分析を行った. さらにアンケートにおいて音声・視覚情報の付与が児童・幼児にとって役に立ったか主観的な評価を行った. ダンス指導者へのアンケート結果についてもまとめを行った. 最後に, UI システムの実験で収集した児童・幼児の動画から定量的にダンス習得度を評価できるか実験を行った.

5.2.1 UI システムの実験

実験参加者の属性は以下の通りである. 実験参加者の人数は 16 人であった.

- 男女人数：男 4 人, 女 12 人
- 年齢別任数：6 歳 9 人, 8 歳 6 人, 9 歳 1 人, 平均 6.917 歳, 標準偏差 1.165
- ダンス歴有無：ダンス歴有 6 人, ダンス歴無 10 人, 平均 0.396 年, 標準偏差 0.887

表 2　グループ別ダンス振付及び音声・視覚情報付与有無比較表

グループ	1 回目	2 回目
A	ダンス a: 音声・視覚 有	ダンス b: 音声・視覚 無
B	ダンス a: 音声・視覚 無	ダンス b: 音声・視覚 有
C	ダンス b: 音声・視覚 有	ダンス a: 音声・視覚 無
D	ダンス b: 音声・視覚 無	ダンス a: 音声・視覚 有

以下の通り実験参加者をグループに分け実験準備を行った (表 2).

1. 実験参加者を 4 グループ (A, B, C, D) に分ける.
2. ダンスの振付を 2 つ (ダンス a(図 5 1), ダンス b(図 5 2)) 用意する.
3. 1 つのダンスの振付について, 音声・視覚情報を付与した動画 (付与有) で練習するグループと音声・視覚情報を付与しない動画 (付与無) で練習するグループに分ける.

以下の通り実験を行った. 実験参加者一人ずつ表 2 のグループ分けの通りの順番で練習を行った.

1. 練習するダンスを動画で 2 回確認する.
2. UI システムでの練習前にダンス振付を行い, それを撮影する.
3. UI システムを使用して 5 分間ダンス練習を行う.
4. UI システムでの練習後にダンス振付を行い, それを撮影する.

上記を 1 回目, 2 回目のダンスで行い, その後児童・幼児にはアンケートに回答してもらった. アンケートの項目については付録 A に示す.

また, 練習前後で撮影したダンスを指導者に確認し, ダンスの動きとリズムについて評価を行った. 指導者はダンス歴 22 年, 指導歴 16 年の X 氏とダンス歴 13 年指導歴 6 か月の Y 氏に依頼した. また指導者の方にはシステムに関するアンケートを行った (付録 B).

＜ *Wilcoxon* の符号付順位和検定＞

アンケート結果を集計し, 音声・視覚情報の有無により, 以下 4 つの項目について検定を行った. X 氏 Y 氏の評価については平均値を用いた.

1. X 氏 Y 氏が練習前後で評価した「動きがよくできたと思いますか？」の値の差の平均値
2. 児童・幼児が評価した「動きがよくできたと思い

ますか？」の値
3. X 氏 Y 氏が練習前後で評価した「リズムはよくできたと思いますか？」の値の差の平均値
4. 児童・幼児が評価した「リズムはよくできたと思いますか？」の値

上記については, 次の略称を以後使用する. 「1. X 氏 Y 氏 A(平均)」, 「2. 児童・幼児 A」, 「3. X 氏 Y 氏 R(平均)」, 「4. 児童・幼児 R」. 音声・視覚情報の有無による 2 群間で母集団の中央値に差があるかを検定するため, ノンパラメトリック手法である Wilcoxon の符号付順位和検定 (Wilcoxon signed-rank test) を実施した. 本検定はデータが正規分布に従わない場合でも有効であり, 本実験のような状況にも適している. 具体的な手順としては, 各ペアの差 $d_i = x_i - y_i (i = 1, \ldots, N)$ を計算し, 差が 0 でないものを抽出した. その際, 差が 0 のデータについては有効サンプル数 (N) から除外した. その後, 絶対値 $|d_i|$ に対して昇順に順位 (rank) を付け, 元の符号 (正負) を順位に戻した. 正の符号に対応する順位の総和 W^+ および負の符号に対応する順位の総和 W^- を算出した. 検定統計量 W は, これらのうちの小さい方 ($W = min \boxtimes (W^+, W^-)$) を採用し, これを用いて「中央値に差がない」とする帰無仮説の検定を行った. 検定は両側検定で有意水準 $p = 0.05 or 0.01$ にて行った. 統計量 W が N 数に応じて, 表 5 13 の数値のとき, 統計的有意差があると判断する. この統計表は Scipy [27] の scipy.stats.wilcoxon を用いて作成した.

＜音声・視覚情報の付与が有効に働いた群の背景分析＞

また, 今回実験したデータに対して解析を行った. 解析は, 被験者の背景属性および児童・幼児の自己評価項目について, 音声・視覚情報の付与が有効であった群 (Snd_Vis-oriented), 音声・視覚情報の付与のない方が有効であった群 (Non-Snd_Vis-oriented), および両者に差が見られなかった群 (Balanced) の 3 カテゴリに分類し, それぞれの特徴を 5 段階スケールのレーダーチャートにより可視化した. まず, 音声・視覚情報の付与によるダンス習得効果を示す 4 つの評価指標 (音声・視覚情報有無それぞれの「音声・視覚情報有 X 氏 Y 氏 A(平均)」, 「音声・視覚情報有 X 氏 Y 氏 R(平均)」, 「音声・視覚情報無 X 氏 Y 氏 A(平均)」, 「音声・視覚情報無 X 氏 Y 氏 R(平均)」の 4 項目) に基づき, 各被験者に音声・視覚情報有りの総合得点 SV_i と音声・視覚情報無しの総合得点 NSV_i を以下の式により算出した ($i = 1, \ldots, 16$)：

$$SV_i = \sum_j^4 sv_{ij},\ NSV_i = \sum_j^4 nsv_{ij} \tag{21}$$

ここで, sv_{ij} および nsv_{ij} は, それぞれ音声・視覚情報有無の各指標である. 次に, 音声・視覚情報有無の得点差 $D_i = SV_i - NSV_i$ に基づき, 以下の基準カテゴリを付与した：

- $D_i \geq 0.5$: Snd_Vis-oriented 群
- $D_i \leq 0.5$: Non-Snd_Vis-oriented 群
- 上記以外 : Balanced 群

このカテゴリ分類に基づき, 以下の 7 項目を対象に平均値を算出した：

- 背景属性：性別 (gender), 年齢 (age), ダンス歴 (dance history)
- 児童・幼児自己評価項目：
  音声・視覚情報有_動き (snd_vis_Act_child),
  音声・視覚情報有_リズム (snd_vis_Rhy_child),
  音声・視覚情報無_動き (Non-snd_vis_Act_child),
  音声・視覚情報無_リズム (Non-snd_vis_Rhy_child)

上記のうち, 性別・年齢・ダンス歴の 3 項目は尺度が異なるため, 5 段階にスケーリングした. これにより, すべての指標を 5 段階スケールで視覚的に比較可能な形式に統一した. 最終的にカテゴリごとの平均ベクトルをレーダーチャート上にプロットし, 各群の傾向を視覚化した.

＜アンケートによる音声・視覚情報付与手法別の主観的評価分析＞

さらに, アンケートにおいて音声・視覚情報の付与が児童・幼児にとって役に立ったか主観的な評価を行い, 表にまとめた.

＜ダンス指導者へのアンケートまとめ＞

ダンス指導者へ実施したアンケートを表にまとめた.

＜児童・幼児の動画分析による定量的評価実験＞

最後に, UI システム実験で収集した練習前後の児童・幼児のダンス動画をコアエンジンで分析し, 定量的にダンス習得度が評価できるか実験を行った.

5.2.2 UI システムの評価

＜ *Wilcoxon* の符号付順位和検定＞

音声・視覚情報の有無による Wilcoxon の符号付順位和検定結果は表 5 14 の通りである. 4 項目すべてにおいて有意差はみられなかった. すなわち, いずれの評価においても, 音声・視覚情報の付与による学習効果の違いは統計的に有意な変化を示さなかった. この結果は, 音声・視覚情報の有無が主観的な動作評価やリ

表 3　音声・視覚情報の有無による Wilcoxon の符号付順位和検定結果

項目	N	W	p=0.05	p=0.01
1. X 氏 Y 氏 A(平均)	12	37.5	-	-
2. 児童・幼児 A	14	36.5	-	-
3. X 氏 Y 氏 R(平均)	12	28	-	-
4. 児童・幼児 R	9	20	-	-

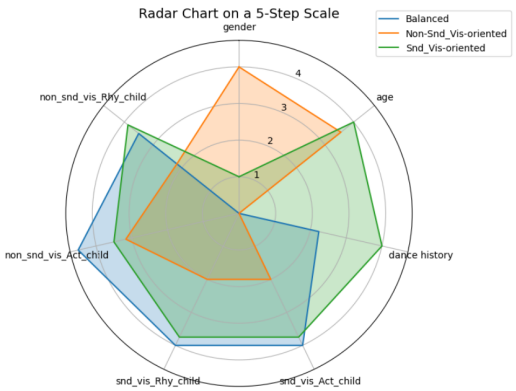

図 3　音声・視覚情報有無による有効群ごとの背景情報レーダーチャート

ズム評価に与える影響が限定的である可能性を示唆している. 特に, 児童・幼児自身の評価 (項目 2 および 4) においても顕著な差が認められなかったことから, 音声・視覚情報の付与が学習者自身の動作感覚の変化に直結しない可能性がある. また, X 氏および Y 氏による評価においても一貫して非有意であったことは, 観察者による外的評価においても同様の傾向が見られることを意味する.

これらの結果は, 今回の実験設定 (回数) では音声・視覚情報の提示が必ずしも学習成果の向上につながるとは限らないこと, あるいは提示方法やタイミングなど, 他の要因が効果に影響している可能性を示唆する.

＜音声・視覚情報の付与が有効に働いた群の背景分析＞

次に, 音声・視覚情報の付与が有効であった群 (Snd_Vis-oriented, $n = 8$), 音声・視覚情報の付与のない方が有効であった群 (Non-Snd_Vis-oriented, $n = 6$), および両者に差が見られなかった群 (Balanced, $n = 2$) の 3 群について, 各群の性別・年齢・ダンス歴と自己評価項目との関連性を分析した (図 3, 表 4). ここで, 表 4 は図 3 の数値を表にまとめたものである. また, 表の性別・年齢・ダンス歴は正規化前の実際の平均値を使用している.

■ Snd_Vis-oriented 群の特徴

Snd_Vis-oriented 群 ($n = 8$) は, 性別平均が 1.13(≒女子中心), 年齢 7.13 歳, ダンス歴 1.34 年と, 他群と比較してダンス経験が長く, 年齢も高めであった. また, 自己評価においても「snd_vis_Act_child」と「snd_vis_Rhy_child」の得点がいずれも 3.75 と高く, 音声・視覚情報提示に対する感受性が高い傾向が見られた. 加えて, 自己評価スコア「non_snd_vis_Act_child」(3.50),「non_snd_vis_Rhy_child」(3.88) も一定以上であり, 全体的に自己評価の高い児童・幼児で構成されていると解釈できる.

■ Non-Snd_Vis-oriented 群の特徴

Non-Snd_Vis-oriented 群 ($n = 6$) は, 性別平均が 1.50(≒男子中心), 年齢 7.00 歳, ダンス歴 0.00 年であり, 未経験の男子児童が主に該当した. 自己評価において,「non-snd_vis_Act_child」は 3.17 と一定の高さを示した一方,「snd_vis_Act_child」と「snd_vis_Rhy_child」はともに 2.00 と低く, 音声・視覚情報提示による効果が出にくい層といえる.

■ Balanced 群の特徴

Balanced 群 ($n = 2$) は, 性別平均 1.00(女子のみ), 年齢 6.00 歳, ダンス歴 0.75 年と, 年齢・経験ともに中間的な位置にある. 自己評価スコアはいずれも高く,「snd_vis_Act_child」と「snd_vis_Rhy_child」は 4.00, また,「non_snd_vis_Act_child」は 4.50,「non_snd_vis_Rhy_child」は 3.50 と高水準にあった. これは, 音声・視覚情報付与いかんによらず, 柔軟に適応できる学習者像を示していると考えられる.

これらの結果は, 児童・幼児において, 性別やダンス経験年数によって音声・視覚情報の有無が学習効果に与える影響が異なる可能性を示唆している. 特に以下のようなことが考えられる.

- ダンス経験のある女子児童には動画への音声・視覚情報の付与が有効
- 未経験かつ男子児童には, 音声・視覚情報の付与の効果は限定的
- 音声・視覚情報の付与いかんに関わらず高得点を示す児童には状況に応じたダンス習得が望ましい.

また, 児童・幼児に「システムをもっとよくするためにこうした方がいいと思うところはありますか？」とアンケートしたところ, Non-Snd_Vis-oriented 群の児童・幼児から,「カメラ表示の記号の〇と棒が混乱し

表 4　音声・視覚情報有無による有効群ごとの背景情報表

	人数	1 女, 2 男	年齢	ダンス歴 (年)	snd_vis		Non-snd_vis	
					Act_child	Rhy_child	Act_child	Rhy_child
Snd_Vis oriented	8	1.13	7.13	1.34	3.75	3.75	3.50	3.88
Non-Snd_Vis oriented	6	1.50	7.00	0.00	2.00	2.00	3.17	2.17
Balanced	2	1.00	6.00	0.75	4.00	4.00	4.50	3.50

表 5　児童・幼児による音声・視覚情報付与手法の主観的評価

	Ave	StDev	Max	Min	Median
見本動画	3.938	0.937	5	3	4
カメラ表示	4.063	1.128	5	2	4.5
オノマトペの音	3.625	1.557	5	1	4
システム	4.125	0.953	5	3	4.5

てしまった.」,「ダンスのことがわからない.」という意見があった. さらに, Balanced 群の児童・幼児からは「先生の動きにもっとあわせられるシステムだとよかった.」との意見があった. 考察として, ダンス未経験の児童・幼児はどこに意識を集中するかのイメージが難しく, 本提案システムはある程度ダンス経験がある方が有効である可能性がある.

＜アンケートによる音声・視覚情報付与手法別の主観的評価分析＞

さらに, 音声・視覚情報の付与が児童・幼児にとって役立ったかの主観的な評価について表 5 にまとめる.

ここで, 表の文言は付録 A の以下質問項目の略称である.

- 見本動画 : 見本動画の文字や記号はダンス練習の役に立ちましたか？
- カメラ表示 : カメラ表示の記号はダンス練習の役に立ちましたか？
- オノマトペの音 : オノマトペの音 (タン) はダンス練習の役に立ちましたか？
- システム : またシステムを使いたいと思いますか？

また, それぞれについて平均値 (Ave), 標準偏差 (StDev), 最大値 (Max), 最小値 (Min), 中央値 (Median) を算出した.

平均値に着目すると,「システム」が最も高く (4.125), 次いで「カメラ表示」(4.063),「見本動画」(3.938),「オノマトペの音」(3.625) の順となっており, 視覚的補助 (カメラ表示, 見本動画) に対する評価が音声的補助 (オノマトペ) よりも高い傾向が見られた. 一方で, 評価のばらつきを示す標準偏差に着目すると,「オノマトペの音」が最も大きく (1.557), 児童・幼児間での評価の個人差が顕著であることが示唆される. 最小値も 1.000 と, 他の項目に比べて顕著に低い. このことから, オノマトペによる提示は一部の児童・幼児にとっては理解や需要が難しい可能性がある. 一方で「見本動画」の標準偏差は 1.0 未満であり,「カメラ表示」も「オノマトペの音」よりも標準偏差が低いことから, 視覚的補助は比較的安定した評価が得られている. また,「システム」は標準偏差が 0.953 であり, 児童・幼児には満足度が高い結果となったため, 練習の習慣化にも使用できる可能性が示唆された.

以上の結果から, 視覚的な情報提示 (「見本動画」,「カメラ表示」) は児童・幼児に対して有効であり, かつ評価のばらつきが小さいことから一貫した学習支援手法として有望であると考えられる. 一方で, オノマトペの音声提示については, 平均値が 4.0 に近い水準を示しながらもばらつきが大きく, 個別の特性や学習スタイルに応じた柔軟な運用が求められる手法であるといえる.

＜ダンス指導者へのアンケートまとめ＞

ダンス指導者へ実施したアンケート (付録 B) を表 5 17 にまとめた. システムの改善点として「ステップ練習において, 足の動きの順序が視覚的に分かる情報からの映像があると望ましい」との意見が寄せられた. 本システムでは前面及び背面からの見本動画のみを使用していたが, 足の動きがダンス習得において重要な構成要素であることを踏まえると, 上方からの映像を取り入れた教材の開発も今後の重要な課題である.

＜児童・幼児の動画分析による定量的評価実験＞

最後に, UI システム実験で収集した練習前後の児童・幼児のダンス動画をコアエンジンで分析した結果を表 5 18 表 5 20 にまとめた.「a_pre」,「a_pst」,「b_pre」,「 b_pst」はそれぞれ「ダンス a の練習前」,「ダンス a の練習後」,「ダンス b の練習前」,「ダンス b の練習後」を表している. 評価については, Snd_Vis-oriented 群, Non-Snd_Vis-oriented, Balanced 群から各 1 名ずつ抽出しコアエンジンの処理にかけ, 抽出した動画フレー

ムと GT である見本動画の抽出フレームにてダイスインデックスを比較した. 結果としては定量的な評価を行うにはいくつかの課題があることが判明した.

一つは, 見本動画との動作の同期に関する問題である. 児童・幼児がある程度見本と同じリズムで動作できた場合には, メトロノーム音および初回の「止め」の動作を基準に同期をはかり, 評価を行うことが可能であった. しかし, 特に初心者においてはリズムに乗ること自体が困難であり, 評価の基準点を特定できないケースが散見された. (今回の動画では, 初回の「止め」の動画フレームにて同期させた) また, 「止め」の姿勢を検出する設計により, 実際には動作が伴っていない場合でも「止め」として誤検出され, 評価が成立しない問題も確認された. 具体的には, 完全に止まっている動画のダイスインデックスが高く, 評価が良いと判断されたしまっていた. これに対して, 見本と類似した座標に手足が位置していることを条件とした「止め」の評価基準の導入など, 今後の改善が求められる.

6. おわりに

6.1 まとめ

本研究では, ヒップホップダンスにおける「止め」の動きに着目し, 動画上の「止め」の瞬間に音声・視覚情報を付与することで, 児童・幼児のダンス習得を支援するシステムを提案した. 提案手法は大きくコアエンジンと UI システムとに大別され, コアエンジンでは動画を入力として音響情報と動画像情報により「止め」の動作を行っている動画フレームを検出し, UI システムでは検出した動画フレームに音声・視覚情報を付与することができた. UI システムを使用した児童・幼児に対する比較実験では, 音声・視覚情報を付与してダンス練習を行った群と音声・視覚情報を付与しないでダンス練習を行った群で統計的有意差が認められるか検定を行った. 音声・視覚情報の有無で統計的な有意差はみられなかったが, 音声・視覚情報の付与を行うことがダンス習得に有効な児童・幼児の背景属性を分析することができた. また, 児童・幼児のシステムへの満足度が高く, 練習に不可欠な継続性の可能性を見いだせたこと, 視覚情報の付与が学習支援として有効であること, 音声情報の付与は児童・幼児の特性を考慮して選択的に使用する必要性があることなどの示唆が得られた.

本研究の主な貢献は以下の 3 点である.

- ダンス動画から「止め」の姿勢を自動的に検出する手法の提案.
- 児童・幼児を対象とした,「止め」の姿勢とタイミングの理解を促すダンス習得支援システムの構築.
- 音声および視覚情報の有無によるダンス習得度の違いに関する実証的評価.

6.2 今後の課題

本提案手法におけるコアエンジン及び UI システムには以下のような課題が残されている.

まず, コアエンジンおよび UI システムの共通の課題として, 評価用サンプル数が少ない点が挙げられる. このため, コアエンジンにおいては, 今回の UI システムで使用した手法以外に, よりロバストに動作を検出できる可能性がある. 音響情報による 2 手法および動画像情報による 3 手法について, サンプル数の増加による精度検証が必要である. 一方, UI システムでは被験者が 16 名と少なく, 統計的検定結果の信頼性に影響を及ぼした可能性がある. したがって, 今度はより多くのサンプルを収集し, 再検証を行う必要がある.

次に, 本研究では撮影された児童・幼児のダンス動画に対して, コアエンジンを用いた定量的評価を行う際に, いくつかの技術的課題があった. 一つは, 見本動画との動作の同期に関する問題, 二つ目は, 実際には動作が伴っていない場合でも「止め」として誤検出され, 評価が成立しない問題である. 今後の改善点として, 見本に類似した動画像上の座標に手足が位置していることを条件とする「止め」の評価基準の導入が求められる.

さらに, ダンス指導者へのアンケートから, システムの改善点の意見が寄せられた. 本システムで取り入れきれなかった部分の開発も今後の重要な課題である.

謝辞 実験データの提供および評価にご協力いただいた DANCE STUDIO NEST の先生方, ならびに実験データ収集にご協力くださった児童・幼児とその保護者の皆様にも, この場を借りて深く御礼申し上げます.

文献

[1] 文部科学省, “小学校学習指導要領 (平成 29 年告示),” https://www.mext.go.jp/content/20230120-mxt_kyoiku02-100002604_01.pdf, March 2017. [アクセス日: 2025-04-12].

[2] 文部科学省, “武道・ダンス必修化,” https://www.mext.go.jp/a_menu/sports/jyujitsu/1330882.htm. [アクセス日: 2025-04-12].

[3] 公益財団法人 日本オリンピック委員会（JOC）, “ブレイキン,” https://www.joc.or.jp/sports/breaking/. [アクセス日: 2025-04-12].

[4] 株式会社 D リーグ, “D.LEAGUE,” https://home.dleague.

co.jp/. [アクセス日: 2025-04-12].

[5] “ヒップホップダンス人口（子ども・青少年）,” https://www.ssf.or.jp/thinktank/sports_life/data/dance_kidsandteens.html, Feb. 2025. [アクセス日: 2025-04-12].

[6] 柳瀬慶子，ヤナセケイコ，“表現運動・ダンス学習におけるリズム系ダンスの「リズムの特徴を捉えて踊る」ということに関する考察: 小学校「サンバのリズム」と中学校・高等学校「ヒップホップのリズム」に着目して,” PhD thesis, Tokoha University, Tokoha University Junior College Repository, 2024.

[7] ダンス・芸能専門 東京ステップス・アーツ，“キレキレのダンスで観客を魅了するための 4 つのコツ and3 つのトレーニング方法,” https://stepsarts.com/column/7208/. [アクセス日: 2025-04-22].

[8] 内山須美子，ウチヤマスミコ，“ストリートダンスのステップを用いた定形型ステップ学習の教育的意義と課題,” 白鴎大学教育学部論集，vol.10，no.1，pp.95–126，2016.

[9] 飯野友里恵，森谷友昭，高橋時市郎，“ストリートダンス動作の分析とダンス指導への応用 (映像表現フォーラム),” 映像情報メディア学会技術報告 35.14 一般社団法人 映像情報メディア学会，pp.49–52 2011.

[10] 亀山有希，カメヤマユウキ，“幼児教育におけるダンス・表現活動の導入に関する研究: リズムダンスを手がかりにして,” 日本体育大学紀要，vol.37，no.2，pp.97–106，2008.

[11] 湯浅理枝，高田康史，“小学校低学年における 「リズム遊び」 の指導法についての一考察-リズムに乗って体幹部を弾ませることに着目して,” 子ども学論集，vol.5，pp.29–40，2019.

[12] 湯浅理枝，“リズム系ダンス授業における児童の着眼点の変容と技能の習得—小学校低学年リズム遊び授業における児童の学習過程に着目して—,” 初等教育カリキュラム研究，vol.10，pp.27–37，2022.

[13] 高田康史，“幼児・児童にもできる簡単ヒップホップダンスに関する実践報告—ipu わくわくリズムダンスの実践を通して—,” PhD thesis，International Pacific University，2015.

[14] 天野海都，三浦健，栫ちか子，“ダンス動画を用いたストリートダンス指導における伝達方法の違いが動作習得過程に及ぼす影響,” スポーツパフォーマンス研究，vol.15，pp.176–185，2023.

[15] 斎藤光，徳久弘樹，中村聡史，小松孝徳，“ダンス動画へのオノマトペ付与によるダンス習得促進手法,” Technical report，情報処理学会，2020.

[16] K. Endo, S. Tsuchida, T. Fukusato, and T. Igarashi, “Automatic dance video segmentation for understanding choreography,” Proceedings of the 9th International Conference on Movement and Computing, pp.1–9, 2024.

[17] 田中佑典，齊藤剛，“モーションキャプチャを用いたダンス上達支援システムの開発,” 第 75 回全国大会講演論文集，vol.2013，no.1，pp.225–226，2013.

[18] 山内，篠本，西脇　亮，絵里子，小野澤，理沙，北原鉄朗，“Kinect とワイヤレスマウスを併用したダンス学習支援システムの試作,” エンタテインメントコンピューティングシンポジウム 2013 論文集，vol.2013，pp.332–338，2013.

[19] 西脇絵里子，小野澤理紗，北原鉄朗，“ユーザーの習熟度に合わせた初心者向けダンス学習支援システム,” 第 76 回全国大会講演論文集，vol.2014，no.1，pp.623–625，2014.

[20] 何毅，谷上明日華，鄭暁潔，彭以琛，吉田匠吾，謝浩然，金井秀明，宮田一乗，“Freedance: 適応型ダンス練習継続支援システム,” インタラクション，pp.••–••，2021.

[21] 戸山恵佑，牛尼剛聡，“ダンサーの振る舞いデータを利用したインタラクティブなダンスチュートリアルの自動生成,” DEIM Forum, F4-3，pp.••–••，2015.

[22] F. Anderson, T. Grossman, J. Matejka, and G. Fitzmaurice, “Youmove: enhancing movement training with an augmented reality mirror,” Proceedings of the 26th annual ACM symposium on User interface software and technology, pp.311–320, 2013.

[23] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” IEEE transactions on pattern analysis and machine intelligence, vol.43, no.1, pp.172–186, 2019.

[24] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “Blazepose: On-device real-time body pose tracking,” arXiv preprint arXiv:2006.10204, pp.••–••, 2020.

[25] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “Vitpose: Simple vision transformer baselines for human pose estimation,” Advances in neural information processing systems, vol.35, pp.38571–38584, 2022.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, pp.••–••, 2020.

[27] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., “Scipy 1.0: fundamental algorithms for scientific computing in python,” Nature methods, vol.17, no.3, pp.261–272, 2020.

Abstract This study proposes a hip-hop dance learning support system for young children, addressing difficulties in understanding timing and posture from videos alone. Focusing on the fundamental “stops” pose, the system combines automatic stops detection using audio-visual analysis with a UI providing onomatopoeic cues and visual feedback. Expert evaluations and Wilcoxon tests showed no overall short-term significance but suggested benefits for experienced girls, with strong user acceptance. These findings indicate the potential of stops detection for early childhood dance education.

Key words dance practice, automatic detection, children, audio support, visual support