Title	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
Author(s)	晴山, 洋人
Citation	
Issue Date	2025-06
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19961
Rights	
Description	Supervisor: 長谷川 忍, 先端科学技術研究科, 修士 (情報科学)

Dance Learning Support System for Lower-Grade Children and Preschoolers Using Audio and Visual Information in Dance Videos

2230029 Hiroto Hareyama

Children usually learn dance step by step, starting with basic movements. These basic movements often include a series of short pauses, known as "stops," where the body temporarily freezes. Children learn the choreography by practicing these "stops" at specific counts in the dance. In most cases, children learn by imitating a teacher's movements. Today, with smartphones being standard, children can also practice by watching dance videos at home.

However, just watching videos is not always enough. Young children often have difficulty understanding the exact timing and posture of these "stops." In real dance lessons, teachers support children's learning by using onomatopoeic sounds (like "tan-tan") and verbal explanations. This step-by-step guidance supports children's understanding of movement.

This study focuses on supporting early elementary school children and preschoolers in learning hip-hop dance. The goal is to help them understand movement better by adding audio and visual information to sample dance videos. The main focus is to improve their recognition of the timing and posture involved in "stop" movements, which can be hard to grasp through imitation alone.

The proposed method consists of two systems. The first is a core engine that automatically detects the video frames where "stop" movements occur. It uses audio and video analysis to do this. The second is a UI system (user interface) that adds sound effects and visual symbols to the detected frames. This information is then shown to the user on screen to support their learning.

In the core engine, "stop" frames are identified based on both audio and video information. For detecting the timing of "stops" from audio, two methods based on the periodic nature of metronome sounds were implemented: one using amplitude features and the other using an autocorrelation function. Skeletal keypoints were extracted using MediaPipe to detect the posture of "stops" in the video. The speed between frames was calculated, and three methods were applied to identify stopping postures: a threshold-based method, peak detection, and k-means clustering. Finally, the system determined the "stop" frames by combining the results of both audio-based and video-based detection.

The UI system takes three inputs: the indices of "stop" frames detected by the core engine, front and back view dance reference videos recorded in sync, and real-time camera footage. Onomatopoeic sounds, text, skeleton lines, and visual symbols are added to the reference videos at each "stop" timing. For the camera view, skeleton lines are also displayed, with their

color changing when a "stop" posture is detected. To support intuitive dance practice, a left-right flip function was implemented for both the reference videos and the camera footage. This UI system enables real-time processing on the CPU.

Next, the two systems were evaluated in experiments. Different methods and parameters were compared for the core engine using the Dice index (Sørensen–Dice coefficient). This helped find the best combination. The results showed that the amplitude features-based method worked best for audio, and the threshold method worked best for video. In the evaluation of video-based methods, the threshold-based approach showed the highest Dice index for all samples. However, the precision values for this method ranged from 0.140 to 0.500, indicating a relatively high rate of false detections, which remains a challenge.

For the UI system evaluation, 16 children participated. They were divided into four groups based on the combination of the presence or absence of audiovisual information and the practice order of two dance routines (Dance a and Dance b): Group A (1st: Dance a with audiovisual information, 2nd: Dance b without it), Group B (1st: Dance a without audiovisual information, 2nd: Dance b with it), Group C (1st: Dance b with audiovisual information, 2nd: Dance a without it), and Group D (1st: Dance b without audiovisual information, 2nd: Dance a with it). Their performances before and after practice were recorded and rated by expert dance instructors. The Wilcoxon signed-rank test was used to analyze whether there were any significant improvements. As a result, no significant differences were found in any of the evaluation items. This result suggests that the presence or absence of audio and visual information may have a limited effect on the evaluation of movement and rhythm. However, when participants were grouped based on their attributes into three categories—Snd_Vis-oriented (where audio-visual support was effective), Non-Snd_Vis-oriented (where support was not effective), and Balanced (no clear difference)—further analysis revealed a trend. Specifically, among older children with dance experience, mainly girls, visualizing the "stop" movements may have helped improve their understanding of the choreography. The Non-Snd_Vis-oriented group consisted mainly of boys with no dance experience. The Balanced group consisted only of girls, with moderate age and dance experience. They showed high self-evaluations and appeared to be flexible learners who adapted well regardless of the presence or absence of audio-visual support.

Subjective feedback from the children was also collected through questionnaires. Many of them said the UI system helped them practice better. This showed high acceptance of the system. In addition, a comparison between the addition of visual information and audio information showed that the addition of visual information resulted in higher evaluation scores with less variability. This trend was consistent with the findings of Saito et al., suggesting that visual support is a promising and consistent method for learning assistance. On the other hand, the feedback for sound cues varied more. This suggests that sound cues may need to be adjusted based on each child’s preferences and learning style.

The contributions of this study can be summarized in three key points:

1. A method to automatically detect "stop" postures in dance videos was proposed.
2. A learning support system was built to help children understand both the timing and posture of "stops."
3. An experimental evaluation was conducted to compare how children learned with or without added audio and visual support.

This study is the first to focus on detecting "stop" movements, which previous research has rarely addressed. It also goes beyond testing audio or visual support separately. It investigates how combining both can help children learn dance more effectively.

Future work will focus on several areas. First, increasing the number of participants will allow for more reliable and generalizable results. Second, we aim to expand the methods for evaluating children's dance performances quantitatively. This will make it easier to measure skill improvement objectively. Third, we plan to enhance the UI system by adding video from different camera angles. This may provide even better support for learning.