Multisensory Learning Framework for Robot Drumming

Andrey Barsky
Claudio Zito
Hiroki Mori
Tetsuya Ogata
Jeremy L. Wyatt

UNIVERSITY OF BIRMINGHAM

Abstract

We present an open-source framework for collecting large-scale, time-synchronised synthetic data across multiple modalities for learning robot manipulation tasks. We demonstrate its use by training a multimodal sensory integration network to generate motion trajectories for robot drumming. We evaluate our system through the quality of its cross-modal retrieval.

Motivation

Integrating **sensory** and **motor** information allows us to learn the causal effects of our actions:

- What will I **perceive** if something is **done?**
- What should I **do** so as to **perceive** something?

Learning invariant representations of sensorimotor actions allows us to predict one from the other.

Training networks of this type requires large-scale synchronised data collection, which is difficult for robot platforms.

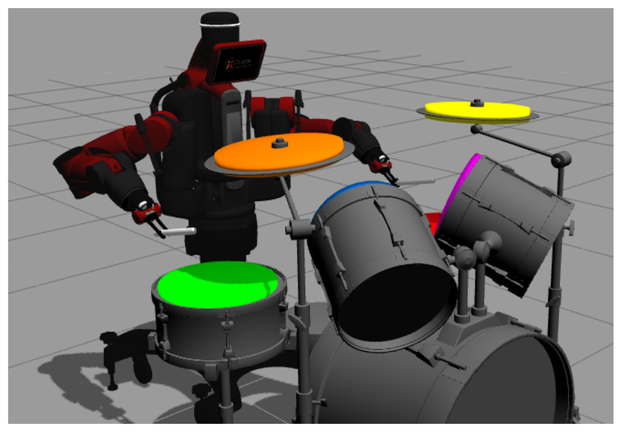
Contribution

We present an open-source framework for quick and efficient simulation and automated recording of synchronised sensory data across multiple modalities.

We then demonstrate its use in the learning of non-linear sensorimotor mappings for a robot drumming task.

Simulation Framework

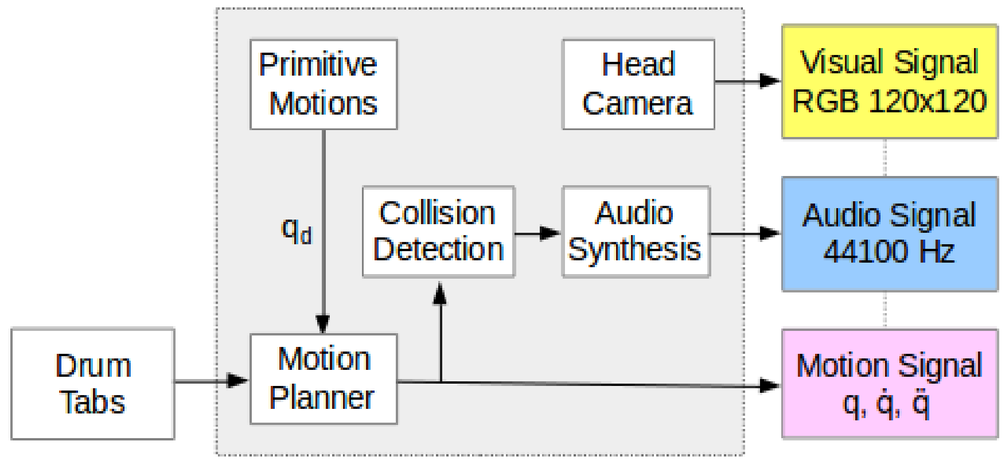
Built in Gazebo 2.0 using Baxter Research Robot SDK and ROS Indigo.



We demonstrate primitive striking motions for each drum and the motion trajectories between them.

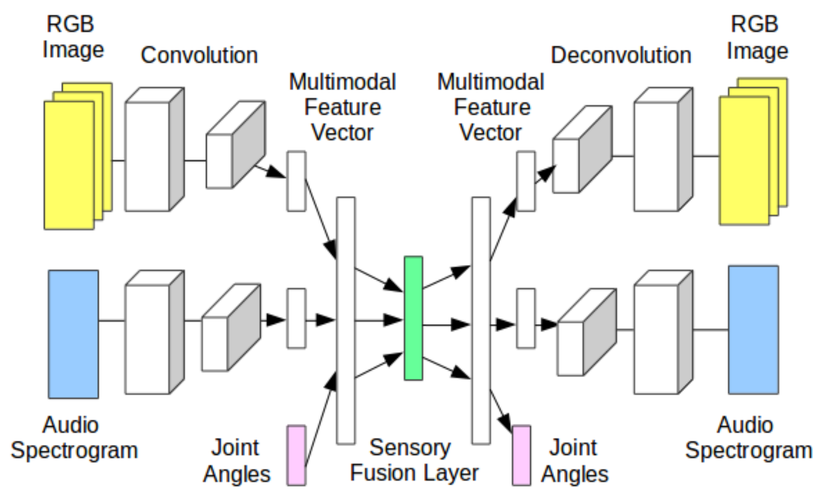
The system accepts desired drum beats and times, and executes a motion trajectory that contacts the target drums at those times.

We synthesise sound to coincide with drum collisions and integrate this with recorded joint angle and camera data.



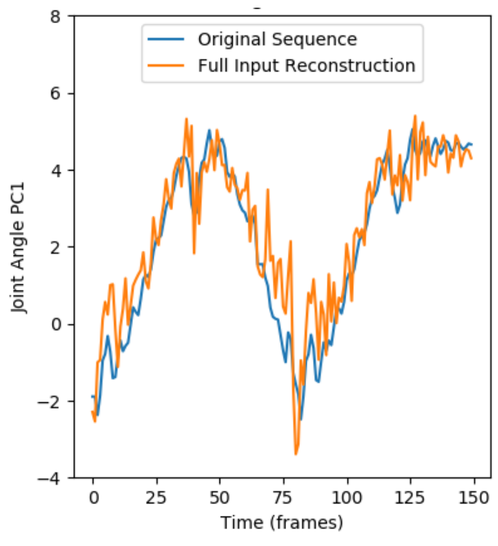
Sensory Fusion Network

We use a multi-channel deep convolutional LSTM denoising autoencoder to extract multimodal features that correspond to invariant representations of drumming movements.



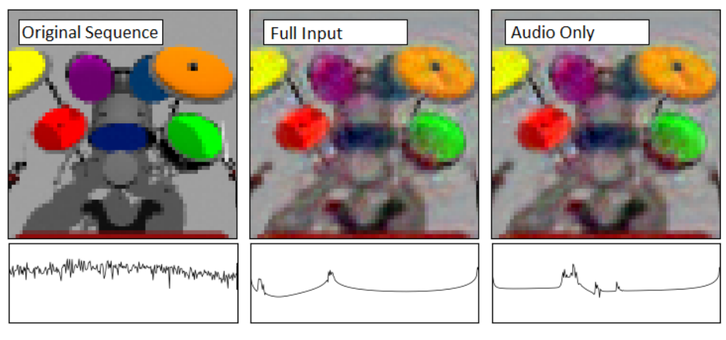
The network is trained to be robust to partial inputs, so that trajectories from one modality may be used to reconstruct those in another.

Cross-Modal Retrieval



Our system can reasonably reproduce multimodal sequences from partial input.

Joint motion from audio is most interesting but requires further hyperparameter tuning for reliable reconstruction.



Summary

Our simulation framework allows cost-effective automated data collection given a specified environment and set of task instructions. Our sensory integration network is able to learn some cross-sensory mappings to generate novel motion from desired audio. A bidirectional RNN architecture may provide better results due to the time dependencies in this task domain.