進捗報告

1 進捗

- STAIR Actions Dataset の一部ダウンロードと整形

2 STAIR Actions Dataset の一部ダウンロードと整形

STAIR Actions Dataset とは, 人の様々なアクションを短い動画にした大規模データセットで, クラウドソーシングによって「誰が」「どこで」「何をしているのか」という日本語キャプション, ラベル付けが 1 本の動画あたり平均 5 つ付いており, 制作者側でラベル付けの結果の検品も行い, 一部のデータを除いている.

現在, 101 種類の動作ラベルがあり, 今回は “drinking”, “eating_meal”, “washing_face”, “gardening”, “fighting” の 5 つのラベルに属している動画の一部をダウンロードし, それぞれの動画から連番の画像と音声を切り出した. 表 1 にそれぞれのデータ数とファイルサイズを示す. 音声の拡張子は “ogg”, 画像は “jpg” に統一した.

3 これからやること

最終目標

- 無音の動画を入力として, その動画にあった音を付ける.
- 日本語キャプションを用いてマルチモーダルな入力から生成させるとどうなるか. (新規性あり?)
- 同様の既存研究は大規模データセットで学習回しているのでそこの削減
- 漫画にも適応していきたい...

やる順番

- 音声波形を入力として動作ラベルを推定する (今ここ)
- 画像 1 枚を入力として元々の音を出力させるように学習
- 動画を入力として元々の音を出力させるように学習
- 日本語キャプションを入力として動作ラベルを推定する
- マルチモーダルな入力から音を出力

生成の過程で一番の課題は, 各データの fps の違いや再生時間の差. 動画のフレームと出力される音のサンプル数をそろえないといけないこと. Auto Foley では事前に ffmpeg を用いてフレーム補間を行って動画の fps を上げて 190 fps で揃えている. 図 1 に Auto Foley のネットワーク図を示す.

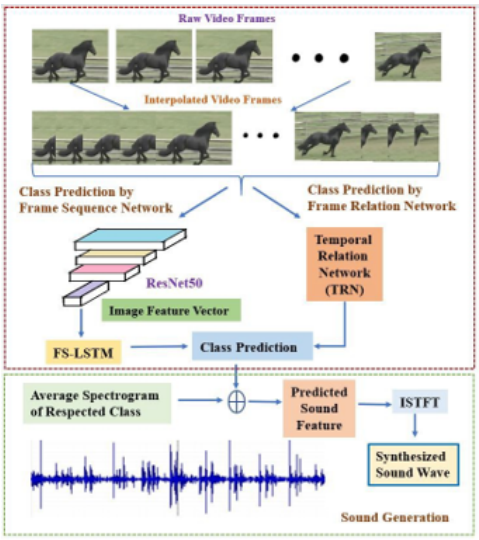

図 1: Auto Foley 概要

表 1: データ数

label	drinking	eating_meal	washing_face	gardening	fighting
動画 / 音声 数	510	441	975	142	272
連番画像 数	97286	81258	162464	23392	47640
データサイズ	33.6 GB	28.0 GB	51.1 GB	9.0 GB	9.8 GB