

行動認識およびマルチメディアデータの自動生成を目的とした 3D モデルアクション動画データセットの構築に関する立案

1 はじめに

人物の行動認識は、与えられた動画中に映る人物が何の動作を行っているのか分類するタスクであり、動画解析における主要な研究テーマの一つとして盛んに研究されている。そして、この応用研究として白黒映像に色情報を付与する自動着色や、映像を滑らかにするフレーム補間、動画に対する音声の自動生成などの生成タスクがある。

近年、アニメやゲームといった動画を扱うコンテンツ業界においても、データの複雑化・大規模化に伴って機械学習を用いた開発・創作支援が取り入れられている。例えば、アニメ業界では中制作業を機械学習によって自動化したり¹、ゲーム業界ではキャラクタのセリフや音声から感情を推定して3Dモデルの表情に適応させる²などの事例がある。

しかし、著作権などの観点から従来の動画解析を目的とした一般に公開されているデータセットは実空間を対象としたものがほとんどであるのが現状である。

そこで、行動認識およびマルチメディアデータの自動生成を目的とした3Dモデルアクション動画データセットの構築に関する立案を行う。

2 Unity, ユニティちゃん

Unity (Unity3D) とは、Unity Technologies (日本法人はユニティ・テクノロジーズ・ジャパン株式会社) が開発・販売している、ゲームエンジンであり、アニメやゲームなど、幅広い業界で用いられている。

ユニティちゃんとは、ユニティ・テクノロジーズ・ジャパン株式会社が提供する開発者のためのオリジナルキャラクターであり、キャラクターを自由に設定できるように利用規約に準じる形でアセット (素材) として無料配布している。

3 GameSynth

GameSynth³ とは、tsugi 社によって開発された最新のプロシージャルサウンド合成技術を駆使してあらゆるジャンルのゲーム効果音を作成できる、サウンドデザインツールで、効果音バリエーションの自動量産機能など、ゲーム開発・映像制作において役立つ機能も含まれている。動画 [1] に GameSynth を用いた足音制作例を示す。

4 データセット制作フロー案

4.1 動作ラベルの選定

「歩いている、走っている、攻撃 (剣)、攻撃 (銃)、魔法 (火)、ギターを弾いている、ピアノを弾いている、階段を昇り降りしている、...」といったアセット (素材) との兼ね合いで実現できそうなものを選定する。

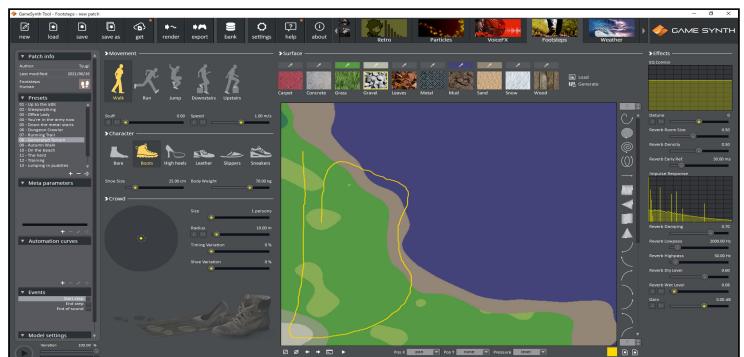
4.2 シーンの作成

各動作ラベルについて、図 1 に示すようなシーンを Unity 上で作成する。次に音声素材を GameSynth 等を用いて制作する。動作ラベルに対応するアニメーションはユニティちゃんアセットに含まれているものや、Mixamo⁴、Unity のアセットストアなどから集める。その他、動作ラベルが「ピアノを弾いている」におけるピアノなど、必要なアセットも Unity のアセットストアなどから集める。(この時、各アセットのライセンスには注意しなければならない。)

4.3 シーンの録画

Unity 内の Record 機能を用いて t 秒間の動画を固定 FPS、固定サンプリングレートで撮影する。一回の撮影には必ず動作ラベルを表すアクションを起こさせるようにし、以下のような情報を毎フレームごとに抽出可能である。ランダマイズするパラメータを指定することで、これを $interval$ 秒おきに録画することでデータの大量作成が可能である。

- 距離、方位角 (azimuth), 仰角 (elevation) (図 2 参照)
- ユニティちゃんのステートが idol か動作ラベルか
- 音イベント検知



動画 [1] GameSynth を用いた足音制作例

¹<https://cacani.sg/>

²<https://cedec.cesa.or.jp/2020/session/detail/s5e58c8811bc98.html>

³<http://tsugi-studio.com/>

⁴<https://www.mixamo.com/>

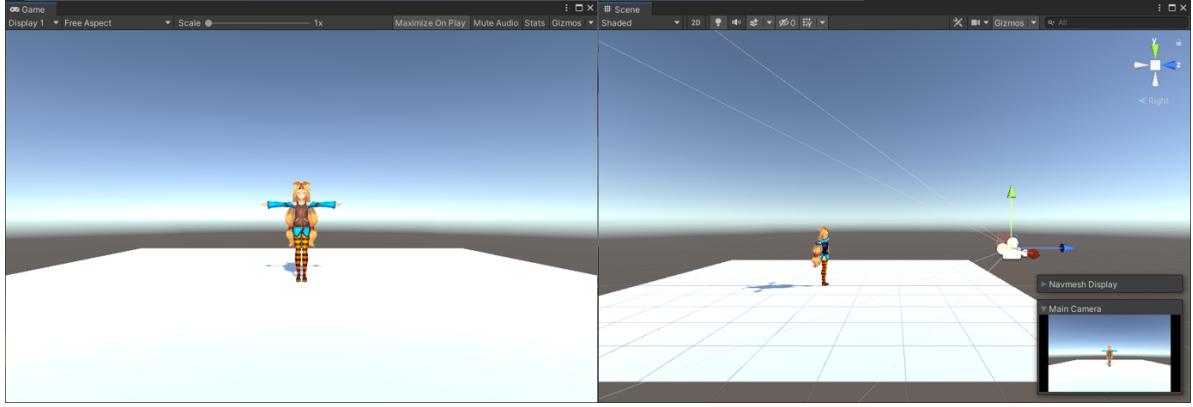


図 1: Unity シーン 概要

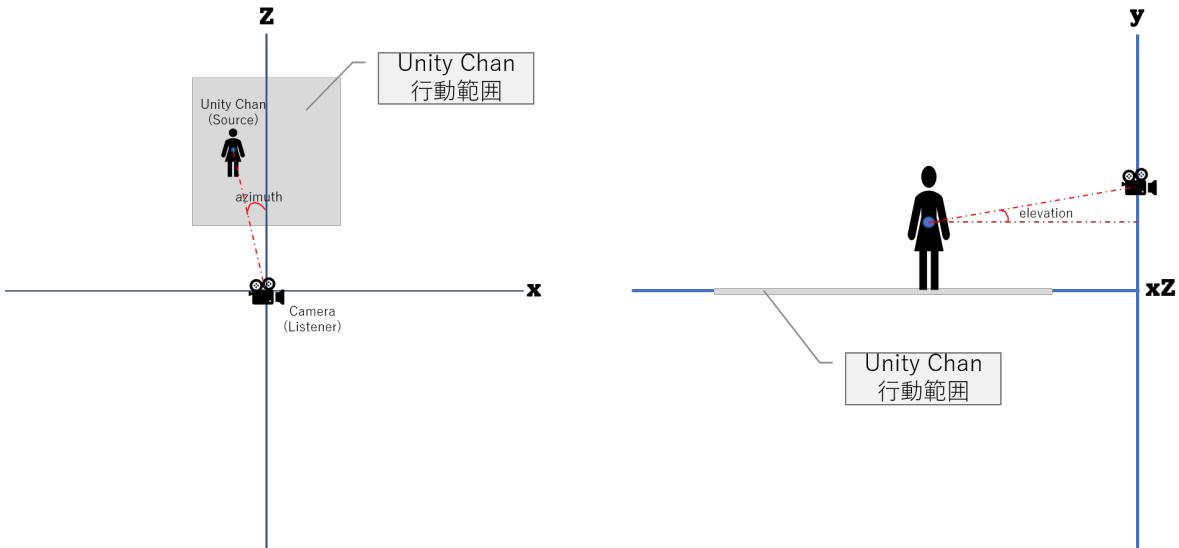


図 2: Unity シーン 座標系

図 3 に出力されるメタ情報の例を示す。

frame	distance	azimuth	elevation	state	audio_event
0	1.71729	-45	14	idle	
1	2.15534	-43	14.3	walk_default	walk_default
2	0.01812	-42.5	14.2	walk_default	
3	0.54333	-39	13.6	walk_default	walk_default
4	1.3783	-23	13.2	walk_default	
5	2.11827	-19	12.6	walk_default	walk_default
6	0.58305	-10	11.6	walk_default	
7	0.25785	-2	11.5	walk_default	walk_default
8	1.66969	3	11.3	walk_default	
9	2.59832	14	11.4	walk_default	walk_default
10	0.5851	16	11.2	idle	
11	0.50678	18	11	idle	
12	2.00431	29	10.4	idle	
13	1.57715	34	9.8	walk_default	walk_default

図 3: 出力メタ情報例 (理想)

4.4 データセットの特徴

- ユニティちゃんを音源、カメラをリスナーとする 3D 音響空間であることから、正確な相対位置をフレームごとに取得できることによって音源の位置推定タスクに有用であってほしい。
- 3D 音響空間であるから、左から鳴った音は左側から聞こえ、遠くで鳴った音ほど小さく聞こえる。
- 3D 音響空間のポストエフェクト（リバーブやエコー）も任意に指定可能。
- 実動画と比べてノイズが入らない。
- 同じ fps、フレーム数、サンプリングレートで出力されるので前処理が楽。エラーデータの削減。

動画 [2] (上) に動作ラベルが「歩いている」のデータ例を示す。現状、動作ラベルが「歩いている」の場合は行動範囲内に目的地がランダムに設定され、これを目標として移動と停止を繰り返すように実装している。将来的には、例えば床のオブジェクトを変えて、床の材質に合わせた足音の変更が可能である。

5 これからやりたいこと

最終目標

- 時系列ごとの行動推定 (ユニティちゃんのステートをフレームごとに推定)
- 無音の動画を入力として、その動画にあった音を付ける。
- 音源位置を考慮した音声の生成 (左で鳴っている音は左側から生成されてほしい。遠くの音ほど小さく生成されてほしい。)

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
- 距離、方位角、仰角を認識できるような映像のエンコーダーを学習 (音源位置推定タスク)
- 音源位置を考慮した音声の生成タスクに取り組む

発展

- 音の生成だけではなく、例えば映像からキャラの表情だけをマスクして、表情を推定・生成するようなことも可能にしていきたい。(できるとは思っていない)
- アプリケーションとして公開できたらうれしい(録画機能がエディタ上でしか動かないらしいので不明瞭)

サブ目標

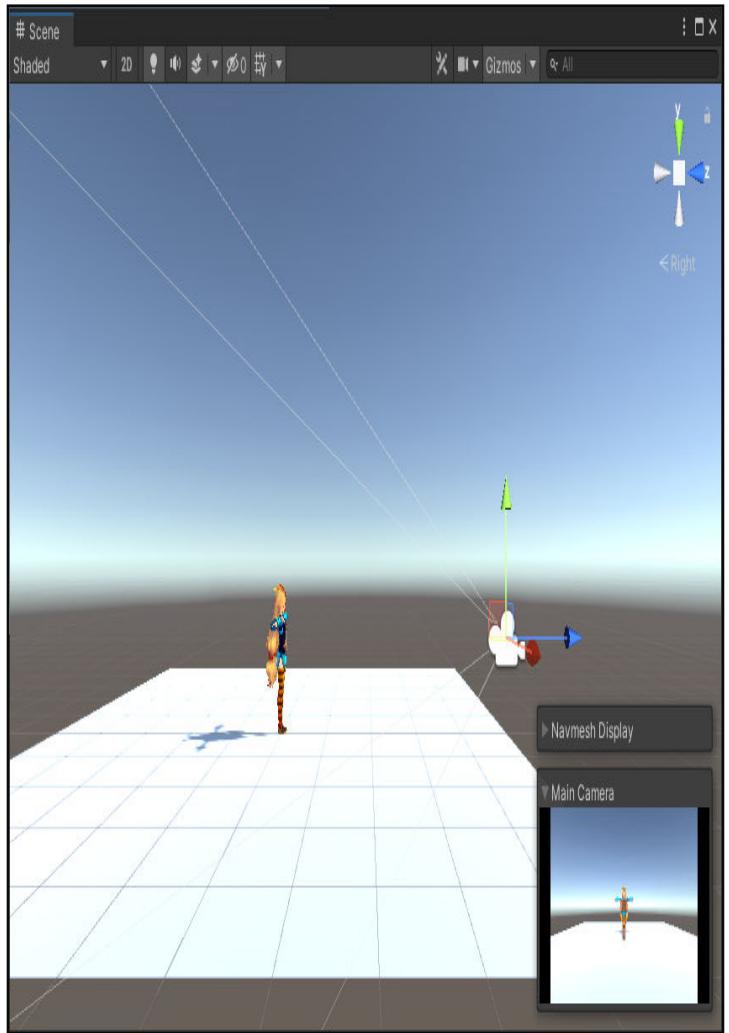
データセットをいくつかの動作ラベルで作成し、音源位置推定を行う。音源位置推定が上手くいきそうになかったら早めに研究テーマを変える。(M1の間に判断) 音源位置を考慮した音声の生成が上手くいくかはその後考える。

修論で聞かれそうなこと

- 新規性 => データセット自体の新規性、(音源位置を考慮した音声の生成)
- こんなものが何に有用か => 創作支援の一点張り

現状

メタ情報がちゃんと出力できるか、データセットの大量生成は可能か、実装できていないので不明瞭。ここが不可能なら早々に打ち切ります。また、そもそもこのような研究が修論として不適切・または不可能に近いと思われるのであれば岡田先生・森先生の判断でストップかけてほしいです。



動画 [2] 「歩いている」データ例