原著論文 Original Paper

ダンスの上手い人のマイニング的な分析

大北 剛 Tsuyoshi Okita
九州工業大学 Kyushu Institute of Technology
tsuyoshi.okita@gmail.com

井上 創造 Sozo Inoue
九州工業大学 Kyushu Institute of Technology
sozo@mns.kyutech.ac.jp

keywords: 行動認識, モービルコンピューティング, ユビキタスコンピューティング, 深層学習, ポーズ推定

Summary

IoT において,「歩く」,「立ち上がる」などの言語による行動のラベルを目的としたセンサからの行動認識を可視化する技術は,「歩く」,「立ち上がる」という言語による行動のラベルを目的とした映像からの行動認識との技術の融合を意味する．これは一転して,「歩く」,「立ち上がる」などの言語による行動のラベルのバイアスを排除する新たな行動認識の形を提案し，新たなマイニングのモデルを提案する. ダンスの上手い人と下手な人のどこが具体的に異なるかをセンサと映像からのマルチモダルな行動認識から探るプラットフォームの構築を報告する.

1. ま え が き

ダンスのような動作を伴う行動を複数の人間で比較する際にまずコーチなどの人間がいれば, ビデオを見て比較が可能となる. また, 三軸加速度などのセンサを用いてよい場合には, 何らかの動作を行なった場合のセンサ値を人間が比較箇所を特定して比較する方法が考えられる. 本論文においては, これを自動で行うにはどうすればよいかという問題を考えたい. まず, ダンスのような動作をビデオ、センサに記録する必要がある. そこで, 本論文ではセンサとビデオを記録するシステムを前半で開発する. 次に, それらの記録した情報をどう解析して人間が行うような解析に繋げるかの考察を行う.

これらの解析を行う動機の一つは行動認識の研究である. 人間の行動認識は, 大きくセンサベースの認識 [6, 5, 8, 7] とビデオベースの認識 (コンピュータビジョンではトラッキングともいわれる) のやり方に分類できる. このいずれのやり方においても, 各々のやり方で自然言語という形で表現された「歩く」「立ち上がる」などの行動へと分類する. この行動認識のやり方を用いると, A 氏が行ったダンスのある動作と B 氏が行った同じ動作を比較した場合に, これらが近い動作をしているか, かなり違う動作をしているかを判断することは比較的容易である. しかし, それらの動作が上手いか上手くないかという判断は極端に難しい. そこで, 設定を容易にするため, A 氏を上手い人と想定し, B 氏が A 氏の動作を真似る場合に, B 氏は A 氏の動作を上手く真似ているか否かという類似する問題へと摩り替えたい. この問題において, それらの動作が上手いか上手くないかという判断は可能となる.

本論文の貢献は以下の通りである.

- ビデオとセンサ信号を同一プラットフォームで収集するプラットフォームの開発,
- 取得したビデオを解析する仕組みとしてポーズ推定による骨組シークエンスを利用する方法の提案,
- ビデオとセンサ信号というマルチモダルなデータにおいてイミテーション学習の成功度を測定する方法として分散表現の累積和を用いる方法の提案

2. センサ/ビデオ取得システムの概要

2·1 ビデオ信号とセンサデータを収集するシステムの構築

まず, ビデオ信号とセンサデータを収集する必要があり, これは python ベースのシステムを開発した. 外観を図 2 に示す. ダンスを行なう人間側はスマートフォンを両手首 (2 台), 両足首 (2 台), 胸部 (1 台) の計 5 台装着すると想定し (図 1 に示す), これらのスマートフォンの三

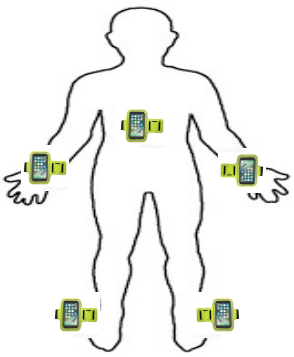

図 1 本論文において用いた 5 つのスマホの位置を示す.

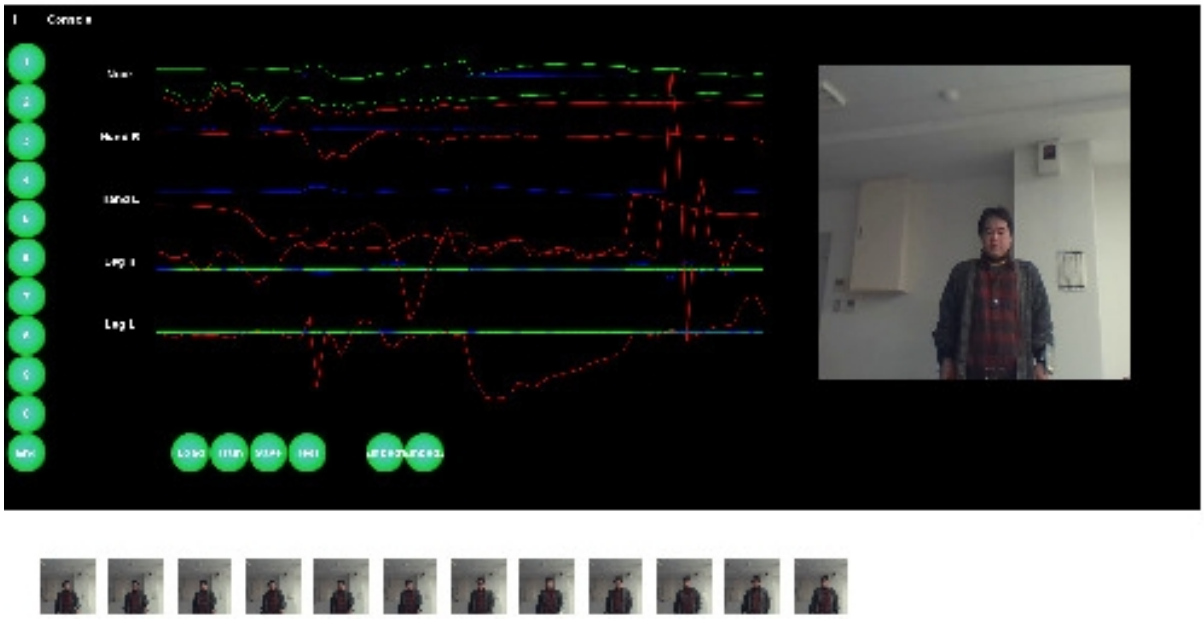

図 2 センサ行動認識においては, 高次元のセンサの読取り値が時系列のシークエンスとして入力され, それに対応する動作を出力とする.

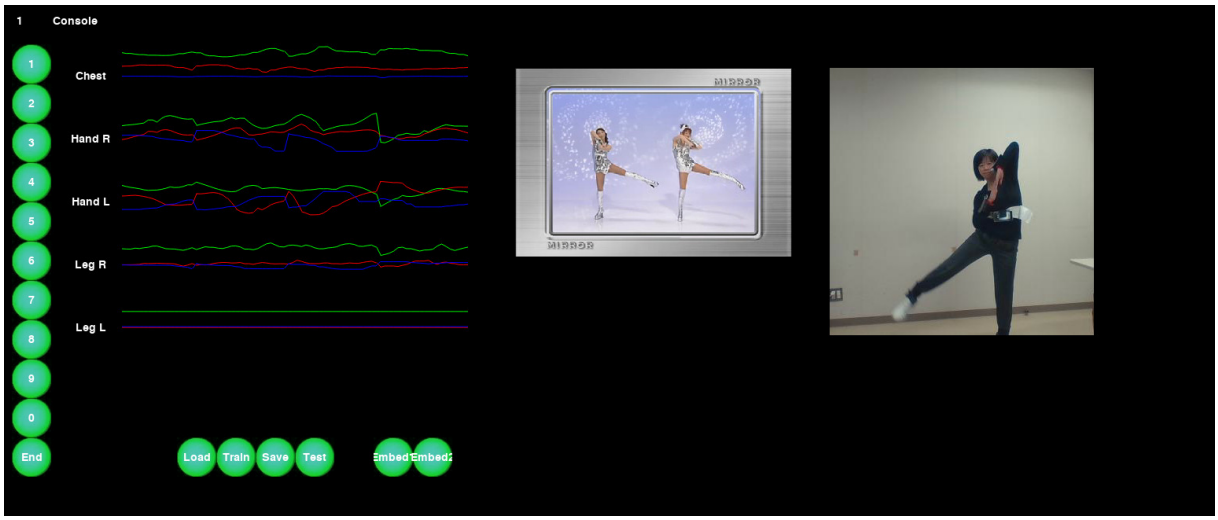

図 3 取得モードを示す. 中央に手本となるビデオ映像を流す.

軸加速度センサの値を本システムで記録する. 三軸加速度センサからの信号は UDP 経由で別々のチャネルで送信させ, 本システム側でその信号を受信する. 本システムはノートパソコン上に実装し, このためビデオ信号はウェブカメラから来ることを想定し, これを本システムで記録する.

内部構成は, センサ処理部ビデオ処理部からなる. センサ処理部においては, 時系列のセンサデータを取得し, 時系列に表示する機能を備える. ビデオ処理部においては, ウェブカメラからのデータを opencv を用いて取得して, システム画面に表示する機能を備える.

取得モード, 再生モード, 解析モードが存在する. 取得モードは図 4 に示す. このモードにおいては, センサデータとビデオ信号を記録する処理を行なう. 振付つきのダンスの取得モードにおいては, 手本となるビデオ画像を音声を伴って流し, これにより被験者が踊りやすくしている.

再生モードにおいては, 記録した信号を再生する. このモードにおいては, 手本となるビデオ画像は省略し, 記録したビデオ映像とセンサデータを再生することを可能としている.

解析モードにおいては, 本論文において述べるような骨組対骨組のアラインメント, 骨組のシークエンスの出力などの機能を持たせている.

2·2 骨組対骨組のアラインメント

取得したデータから骨組を得るためにビデオ映像を画像シークエンスに落とし, Openpose[1] を利用して各画像に対して骨組および 45 次元の骨組ベクトルの座標を得た. 図 5 はサムネイルの例を示し, また, 図 6 は骨組ベク

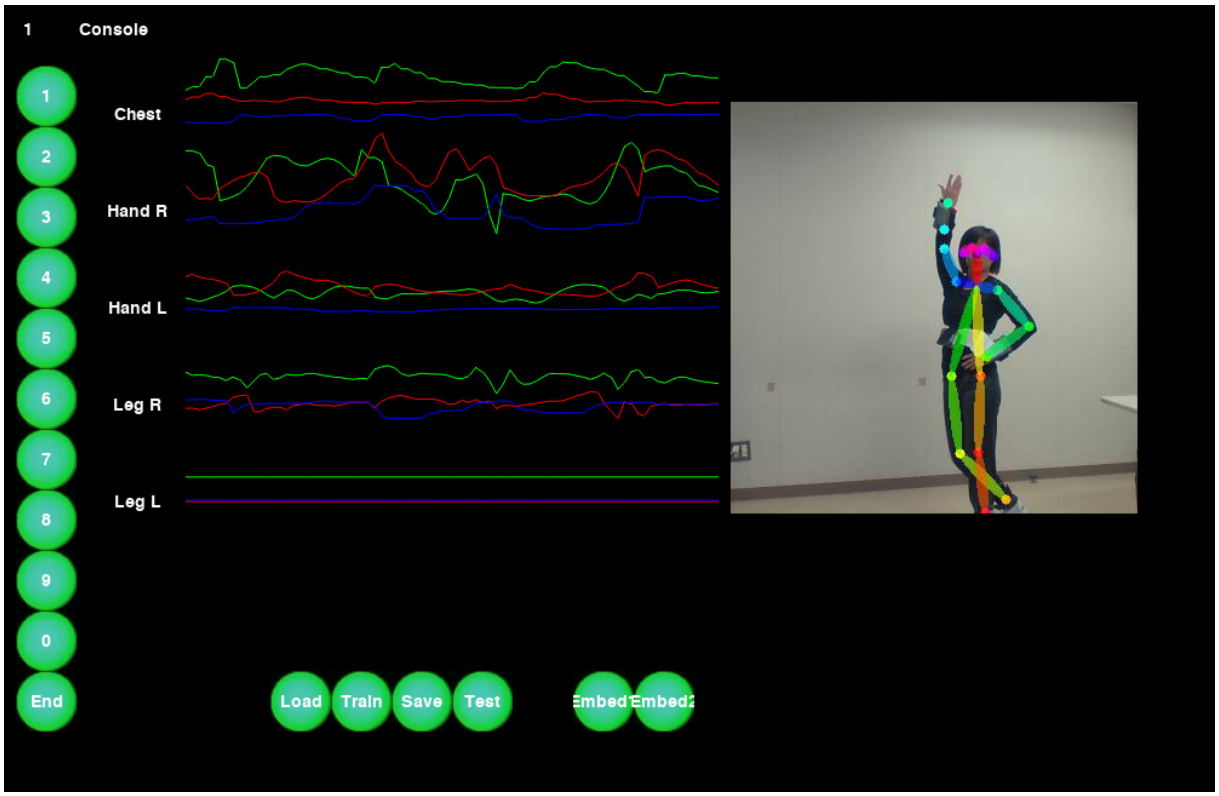

図 4 再生モードを示す. 左側には取得したセンサデータ, 右側には骨組解析をオーバーラップさせたビデオ映像を流す.

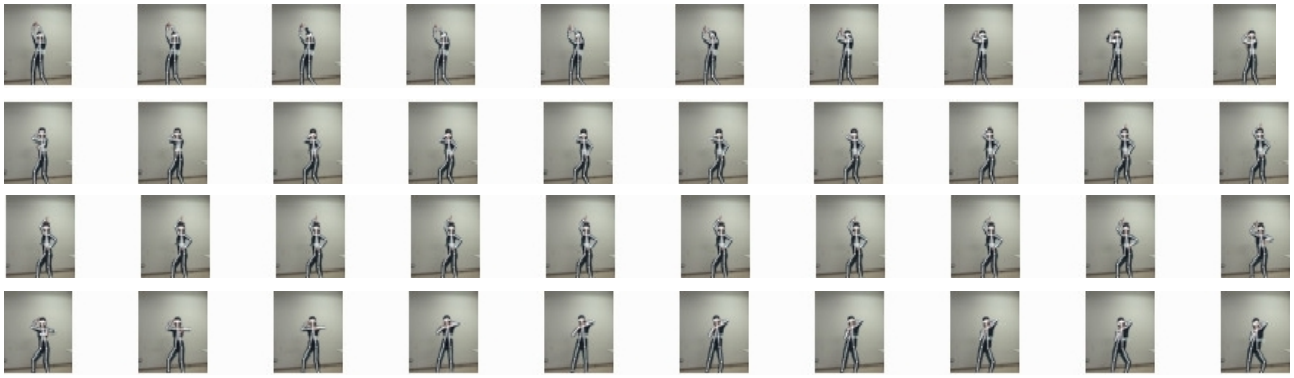

図 5 画像シークエンスにおいて, 各々の画像に Openpose を適用して, 骨組をそれぞれ得ている. この後さらにこれらの画像シークエンスをビデオ映像に変換する.

トルを表示したものを示す. これらの画像シークエンスをビデオ映像に変換した. ビデオ映像に変換した後はビデオ映像は骨組のみとなる.

3. ダンスの解析

ダンスを解析することは, 本論文においてはセンサデータおよび骨組ベクトルを解析することと帰着させる. 本論文においては, 手法を紹介して簡単な分析を行ない, 実行可能性を吟味するに留める. 自由なダンス曲で個人技の巧さを分析する形のものは考えず、固定したダンス曲、たとえばピンクレディの UFO の振付け, を上手い人と上手くなろうと努力している人の違いは何かを分析する形のものを考える. なお, 自由なダンス曲で個人技の巧さを分析する形においてはこのような形で上手い人のダンスの分析をする場合には, 使っているボキャブラリの数などを比較することができるはずである. この場合, 本論文の解析方法そのものでは対処はできず若干の拡張を必要とする.

主要なプロトコルは以下の 5 つとした.

(1) 上手い被験者 (以後, A と呼ぶ) がセンサをつけた状態でダンスを行ない, 同時にビデオ撮影も行なう. その後, 上手くなろうとする被験者 (以後, B と呼ぶ) がセンサをつけた状態でダンスを行ない, 同時にビデオ撮影も行なう.

(2) 振付けにおいてシーンを設定し, ビデオをシーン毎に区切るラベリングを行なう. この後, それぞれのシーンの開始を A と B においてアラインメントを

図 6 骨組シークエンスのみとする.

行なう. さらに, それぞれのシーンにおいて, A と B の骨組をアラインメントさせる.

(3) B が外して振付と全く違う動作を行なうことが頻出することが予想されるが, この場合には, アラインメントはあえてしないようにする.

(4) センサデータに関しては, ビデオにおけるシーンの区切りと同期する時点をそれぞれのシーンの開始点と定義する.

(5) この同期する点は, ビデオ信号とセンサデータの遅れを反映させた形のものとなり, しかし, すべての同期点において一定の時間の遅れと考える.

項目の 1 番目であるが, 我々の分析の目的を細分化して, イミテーション学習を行うコンテクストとする. つまり, 上手い被験者 (以後, A と呼ぶ) がセンサをつけた状態でダンスを行ない, 上手くなろうとする被験者 (以後, B と呼ぶ) がセンサをつけた状態でダンスを行う. このようにして記録したビデオとセンサデータを比較するという方法を取る.

前述したようにビデオは, ポーズ分析 [1] を行ない, 抽象化されて骨組となる. この抽象化により, 該当する部位, たとえば A の右手と B の右手, を比較することにより行なえることとなる.*1 また, 今回はこれ以外の比較は行なわず, 骨組のみで比較を行なう. 比較を行なうためには, シーン毎に A と B の骨組がアラインメントされている必要がある. このため, 項目の 2 番目であるが, A と B のアラインメントした骨組を比較するためには事前設定として, 振付けのシーンを設定して, ビデオ先導でシーンを区切っておく. つまり, それぞれのシーンの開始を A と B においてアラインメントを行なう. さらに, それぞれのシーンにおいて, A と B の骨組をアラインメントさせる.

項目の 3 番目は, A と B の距離を求めるのが困難な場合である. B が外して振付と全く違う動作を行なえば, A と B の距離の解釈が著しく困難となる. この場合には, アラインメントはあえてしないようにする. また, この期間は B のイミテーションを行なうスコアはゼロと考える. 一方, アラインメントしている場合には, A と B の骨組の差を比較して, そのスコアを骨組の要素ごとに行なう.

項目の 4 番目は, センサデータに関しては, ビデオにおけるシーンの区切りと同期する時点をそれぞれのシーンの開始点と定義するというものである.

項目の 5 番目は, センサデータとビデオのシーンとの間には遅れが存在する点で, これが無視できない. これは, 我々のシステムにおいて, センサデータは常にビデオ信号より若干遅れて到着し, この記録も若干遅れることによる. そこで, センサデータとビデオのシーンとの間の遅れはすべての同期点において一定の時間の遅れと考えるというものである.

4. 行動認識との比較

センサベースの行動認識は, 入力を高次元のセンサ信号とし, 出力を人間の行動とする. 教師あり行動認識として機械学習を用いる方法を本論文では論ずるため, ここでは教師あり行動認識のみを考慮する. 教師あり行動認識の場合, 行動クラス $\mathcal{Y}$ はトレーニング集合に有限個の行動として定義され, たとえば, $Y = \{$ 立ち上がる, 歩く, ジョギングする, $\ldots\}$ などとなる. 一方, センサデータ $\mathcal{S}$ は採取に用いるセンサの種類と数に依存し, センサデータの次元は $n(=\{1,\ldots,N\})$ となる. 各々のセンサは時系列のデータで構成され, k 番目のセンサ $(1 \leq k \leq N)$ に対する時刻を $t(=\{1,\ldots,t_k\})$ と定義すると, $s^{(k)} = (s_1^k, s_2^k, \ldots, s_{t_k}^k)$ と表現できる. ここで人間の行動を記述するのは「立ち上がる」「歩く」「ジョギングする」などの自然言語である.

ビデオベースの行動認識は, 入力をビデオ信号とし, 出力を人間の行動とする. 本論文で行なうシーンの解析においては, 上手い人の行動をいかに上手く真似たか (イミ

*1 一方, この単純化は多くの現実的な因子を失うことは確かである. たとえば, 骨組のみをビデオにした場合, ダンスが上手いか下手かの判断が極端に難しくなる. 肉付けしても未だ難しく, テクスチャマッピングまでやらないと判断が難しくなるように思える.

テーションしたか) という因子を解析することになる. ここでも人間の行動を記述するのは「立ち上がる」「歩く」「ジョギングする」などの自然言語となる.

さて, 本論文における解析は, センサベースの行動認識やビデオベースの行動認識で用いるラベルの部分が大きく異なる. 一つ目, 我々の目的において, シーンに対してラベルづけは行なう. シーンはたとえば,「リズムキープ」「フォーコーナー」「腕をのばしたまま振る」「腕を止める」「ポーズ」「ウォーク」[*2]などの他, どちらかというとこれらが同じシーンにおいて A と B の動きを比較するための前準備に使いたい. マイニングできるためにはこれらの分類は必要となるがこれが本質ではない. 二つ目, これらの分類は上手く真似たかの指標とはなりえないため, 上手く真似たかの指標となるものを得る必要がある. 上手く真似たかの指標は, A と B のセンサデータとビデオ映像を比較する折に得られる. 三つ目, B の動作はエキスパートである A の動作とは似ても似つかぬ動作を行なっている可能性があり, これらの動作をしている箇所は比較の対象から外すのが無難であろうと思われる. このために, 同一のシーンにおいて, A と B の分散表現の距離の累積和を比較して, たとえばコサイン距離の累積和が閾値より大きければ対象から外すというやり方を選択した. 四つ目, センサデータの方も分散表現の距離の累積和の比較というやり方を選択した. 五つ目, 累積和の比較と言ったが, 振付の場合, 音楽と同期させるため, 同期ポイントはより明確なはずで同期した時刻における比較を行ない, かつ, 同期していない時刻における比較を行なわないのがよいかと思われる.

4·1 分散表現の距離の累積和の比較

ビデオ側においてもセンサ側においても分散表現を用いて比較するため, 以下のような方法を用いた. シークエンシャルな性質を考慮すると, センサ信号とビデオ信号/骨組の分散表現は等価である. したがって, センサ信号をエンコーダーデコーダ [2, 4, 9] を用いて翻訳して構築した分散表現を取り出して用いる方法である. そして, それをシーンにおいて A と B の分散表現の距離の累積和を比較するやり方である.

5. 実　　験

以上の動機のもと, センサの時系列シークエンスから行動のシークエンスへエンコーダーデコーダー型 [2, 4] を用いて翻訳することを考える. したがって, デコーダ側は時系列ごとにビデオシークエンスをラベルとして骨組ベクトルを用いる. この設定は率直にはいかないため, 以下のような工夫を行なう.

一つ目, 生のセンサ信号、ビデオシークエンスはいずれもストリームだが、エンコーダーデコーダー型では無限長を扱ってはおらず, 逆にセンテンス長の長い領域では精度が悪くなる. この長さ制限を考慮して、暫定的にストリームをエンコーダーデコーダーいずれの側においても, 100 ユニット以内に留めた. エンコーダー側はセンサの周波数、デコーダー側はビデオの周波数でいずれも異なるため、エンコーダー側の開始時点の時刻、終了時点の時刻に合わせる形でビデオ信号の始点、終点を定める. この制約を満たすように設定すると, エンコーダ長 61 項目, デコーダー長 45 項目となった. 測定に用いたセンサの周波数、ビデオの周波数は異なる被験者においても同じ設定で用いたため、この 61 項目と 45 項目という比率は固定とする.

二つ目, エンコーダ側の 1 項目は 3 軸加速度センサを 5 台装着したことによる 15 次元からなり, デコーダー側の 1 項目はビデオ信号をポーズ分析した結果の 45 次元からなる. つまり, 加速度センサの 15 次元, ポーズ情報の 45 次元を入力と考え, 一方, エンコーダ長, デコーダ長の 61 と 45 という項目数は, センサの周波数とビデオの周波数を修正したものと考えたことになる. なお, 加速度センサの 15 次元 (もしくはポーズ情報の 45 次元) を一般的に扱うことはせず, 凝集型 (agglomerative) なクラスタリングを行ない, ユークリッド距離に関して類似する点を同等と考え, 1000 クラスから 4000 クラスのラベルを貼った.

三つ目, 被験者は 3 人のデータを用いたが, 暫定的に個人差はないものと考え, また, ダンス, 運動などによる差はないものと考えた. 具体的には, (1) 歩く、走るなどの日常動作のデータ, (2) ラジオ体操第 1 のデータ, (3)UFO のデータの 3 種類を用いることにした. これら 3 種のデータは 20,000 対作成し, これをトレーニング, 検証, テストに 20,000 対, 700 対, 660 対にそれぞれ分割した.

この設定により, センサ信号の入力列があれば, 対応するポーズの列が出力されることになるがここではこの機能は使わない.

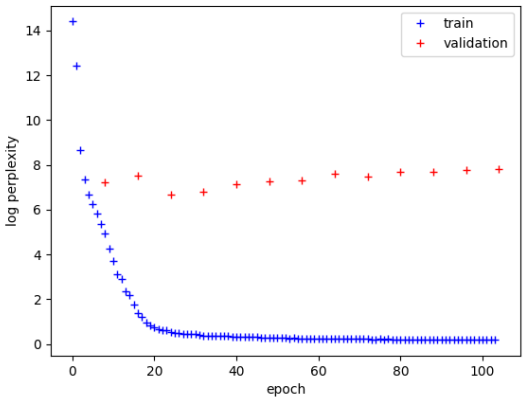

図 7 パープレキシティの推移.

*2 これらの記述は http://www.asakumamasaru.com/entry/waack-dance による.

6. ま　と　め

本論文においては, ダンスの分析を行なうプラットフォームの構築と暫定的な解析を行なった.

問題が難しいため, 今後の話題はいくつも見付かった. 一つ目, 骨組ベクトルの表現を用いたが, 本当にこの表現で巧さの指標となることはもしかすると疑問かもしれない. なぜなら, これを映像にしてみた場合に上手いか下手かの判断がつきにくいことによる. 肉付けして, テクスチャマッピングが必要かもしれない可能性はある. 二つ目, エンコーダデコーダ型の学習器でデコーダ側のセンテンス長に制約をつけることが困難そうなことである. 三つ目, 凝集型のクラスタリングを用いたが, そもそも, クラス数が 4000 程度の場合, 骨組の映像を見た場合にスムーズさがなく, さらなるクラス数が必要そうであることである.

謝　　辞

伊藤雅子さん, 武田紳吾くんには貴重な時間を割いてダンスをしていただき感謝の意を表します.

◇ 参　考　文　献 ◇

Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, In Proceedings of CVPR 2017, 2017.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing (EMNL

井上創造, ウェアラブルセンサを用いたヒューマンセンシング, 知能と情報, 28:6 pp. 170-186, 日本知能情報ファジィ学会, 2016 P 2014). 2014

Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR 2015. 2015

井上創造, ウェアラブルセンサを用いたヒューマンセンシング, 知能と情報, 28:6 pp. 170-186, 日本知能情報ファジィ学会, 2016

Kawaguchi, N., Ogawa, N., Iwasaki, Y., Kaji, K., Terada , T., Murao, K., Inoue, S., Kawahara, Y., Sumi, Y., and Nishio, N. Hasc challenge: gathering large scale human activity corpus for the real-world activity understandings. In Proceedings of the 2nd Augmented Human International Conference, ACM, 27. 2011.

Tsuyoshi Okita, Sozo Inoue. Recognition of Multiple Overlapping Activities Using Compositional CNN-LSTM Model. Ubicomp Poster, Sep, 2017.

Francisco Javier Ordonez, Daniel Roggen. “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. S ensors 16:115, 2016.

Felix Hill, Roi Reichart and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. Computational Linguistics. Vol. 41, No. 4, Pages 665-695. 2015.

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理