*The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018*

# 講義代行ロボットにおける双方向 LSTM を用いたジェスチャ自動生成システムの性能評価

## Performance Evaluation of Automatic Gesture Generation System Using Bi-Directional LSTM on Humanoid Robot

日和 航大[*1] 荒木 健治[*1] 長谷川 大[*2] 芳尾 哲[*1]
Kodai Hiyori Kenji Araki Dai Hasegawa Satoshi Yoshio

[*1] 北海道大学大学院情報科学研究科 Graduate School of Information Science and Technology, Hokkaido University
[*2] 東京工科大学 Tokyo University of Technology

Conventional lecture substitution systems with humanoid robots use pre-defined gestures created by hand. Automatically generating these gestures makes it possible to create gestures without requiring expert knowledge and labor, which is expected to lead to further progress in research on lecture substitution systems. This paper proposes an automatic gesture generation method which is expected to consider the semantic context of an utterance. Our proposed method is implemented using a deep neural network with Bi-Directional LSTM units, applying filters for data correction, and axis conversion.

## 1. まえがき

近年, Softbank 社の Pepper[*1] や NAO[*1] のようなヒューマノイドロボットの発展が目覚ましく, かつては人間が行った仕事やサービスを代行することが可能になっている. また, 橋本ら [橋本 2012] や Xu ら [Xu 2014] によるヒューマノイドロボットを用いた講義代行の研究が行われていて, 今後ヒューマノイドロボットが教育の現場で導入される可能性は十分に考えられる.

現在の講義代行システムでは, ジェスチャは人手で作成されている. 自動化を実現することで, 専門的知識や労力を必要とせずにジェスチャを作成でき, 講義代行システムの開発におけるコストの削減が可能となる. このため, ジェスチャの自動生成化は講義代行システムの開発において, 必要不可欠な要素であると考えられる.

本研究の最終目的は, ジェスチャを自動生成し, 人間の講義よりも質の良い講義代行システムを開発することである. 本稿では, その第一段階として, 時系列データの関係性を考慮できる Bi-Directional LSTM (BLSTM)を使用した RNN により, ジェスチャを自動生成するシステムを Pepper を用いた講義代行システムに用いた際の印象評価実験およびジェスチャの自動生成手法の性能評価実験の結果, 考察について述べる.

## 2. 関連研究

ヒューマノイドロボットが講義を行う研究として, 橋本らの研究が挙げられる. 橋本らは, SAYA [橋本 2006] に小学校の理科の実験を代行させ, 印象評価実験を行った. ジェスチャは実験者が別室から SAYA を遠隔操作することで表現されている.

また, Xu らの研究では, 人手で作成されたポジティブ, ネガティブなジェスチャによって, 講義の印象がどのように変化するかを調査した. ヒューマノイドロボットには NAO を用いている. ポジティブな動作とは, 身振りが大きく活発な動きのことを指し, 一方, ネガティブな動作とは, 身振りが小さく動きも遅いジェスチャと定義される. ポジティブな動作とネガティブな動作で比較実験を行い, ジェスチャの動作による印象の変化を調査した.

一方, 我々は Pepper を用いて人手で作成したジェスチャを行う講義代行システムを開発し, 印象評価実験からジェスチャが講義にどのような印象を与えるかの調査を行なった[日和 2017].

本手法は, ジェスチャの作成方法が, 人手によるものでなく, 発話内容を考慮した自動生成である点で関連研究と異なる. 本稿の目的は, 自動生成されたジェスチャを用いた講義代行システムの印象評価およびジェスチャ自動生成手法の性能評価から, 自動生成されたジェスチャの発話内容に対する妥当性を評価することである.

## 3. 講義代行システムの処理過程

講義代行システムは GUI 開発ツールの Choregraphe[*2] を用いて実装した. 本システムの講義形式はスライドを用いた講義となる. 処理過程を図 1 に示す.

処理過程全体の概要を述べる. まず, 講義開始時に簡単な挨拶を行う. 挨拶を終えた後, スライド1ページから生成された講義内容とジェスチャを読み込む. 講義内容とジェスチャの再生後, 次のページの内容とジェスチャを読み込む. この一連の流れを最後のページまで繰り返すことで講義は終了となる. ジェスチャは 4 章で説明する BLSTM を用いた RNN のネットワークにより自動生成されたものである.

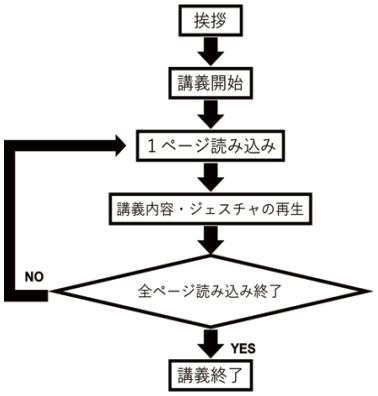


図 1 講義代行システムの処理過程

連絡先: 北海道大学工学部メディアネットワークコース言語メディア学研究室
〒060-0814 札幌市北区北 14 条西 9 丁目
E-mail:kururi@eis.hokudai.ac.jp

*1 http://www.aldebaran-robotics.com/ja/.
*2 http://developer.softbankrobotics.com/jp-ja/downloads/pepper

## 4. ジェスチャ自動生成手法

### 4.1 BLSTM を用いたジェスチャの自動生成

#### (1) データセット

学習に使用するデータセットは，竹内らが作成したモーションキャプチャデータセット [竹内 2017] である．データセット内には音声とそれに伴うジェスチャのペアが含まれる．ジェスチャのモーションデータは，関節の位置と回転角がフレーム毎に記載される時系列データである Bio Vision Hierarchy (BVH) 形式として保存される．データセットの増量と過学習対策のため，元の音声にホワイトノイズを合成した音声とジェスチャのペアを追加し，合計 2,098 センテンス (約 392 分) を学習と検証に用いた．

#### (2) 入力データ

入力データの特徴量として，音声認識において一般的な Mel-Frequency Cepstrum Coefficients (MFCC) を使用した．MFCC は声道特性を表す特徴量である．

図2に示すように，ある時刻に対して，timestep を 0.05 とし前後 30step 分のコンテキストを抽出した MFCC を入力とする．ジェスチャは一瞬の音素よりも，単語や文節の前後のコンテキストにより決定されるため，LSTM による時系列的な連続性の考慮だけでなく，一定範囲の音素の考慮も行う．

正解データとなる，入力データのラベルは時刻 t における各関節の位置と回転角のベクトルとなる．

#### (3) ネットワーク

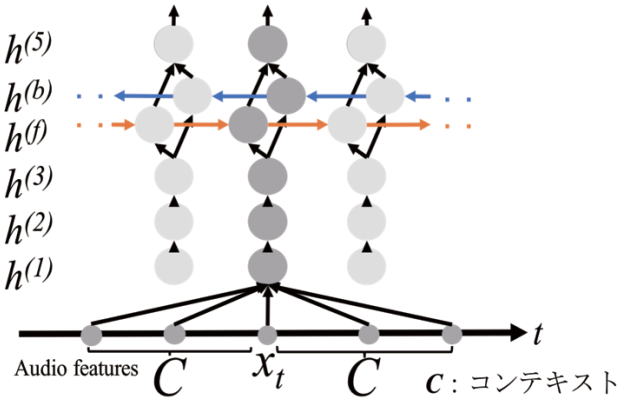


図 2　ネットワークの概略図

本ネットワークは竹内ら [竹内 2017] の手法を参考にしたものである．ネットワークの概略図を図 2 に示す．

BLSTM を用いたネットワークから学習モデルの構築を行う．

ネットワークは全 5 層の隠れ層で構成される．最初の 3 層は同じ構造で，再帰でない全結合層になる．

4 層目は BLSTM 層にあたる．この層により，時系列の前後の関係性を考慮することが可能となる．

5 層目は出力層であり，出力は BVH 形式における足の関節を除いた 43 関節の xyz オイラー回転角のベクトルとなる．

出力結果から平均二乗誤差により誤差 (loss) を算出した．データの学習回数を 500 回にし，学習を行ったところ最終的な loss は 27.4 となった．

### 4.2 フィルタ補正

出力となる回転角の時系列データにおいて，フレーム飛びと見られる箇所の動きを滑らかにするために，移動平均フィルタを用いて出力データの平滑化を行った．このフィルタを用いることで，データの特徴を残したまま滑らかなデータにすることが可能になる．式(1)に示すように，時系列データ$a_1, a_2, \ldots, a_n$に対し，直近 n 個の平均から新しい時系列データ$b_1, b_2, \ldots, b_n$を定める．

$$b_n = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n} \quad (1)$$

CG の棒人間でジェスチャの動きを確認したところ，窓枠 n=3 では依然ジェスチャがぎこちなく，n=9 では滑らかではあるが，値が間引かれたことによるジェスチャの遅延が目立ったため，今回は窓枠を n=5 とした．

### 4.3 座標軸変換

最後に，Pepper の各関節角度と対応するように，フィルタ補正を行ったデータに後処理として座標軸変換を行う．

BVH の座標軸と Pepper の座標軸を一致させて各関節で対応する軸の回転角を代入させる．具体的には BVH の姿勢を，各関節角度の値が 0 の時の Pepper の姿勢 (以後,初期ポーズという) に統一し，関節毎に対応させる．

図 3 に示すように，BVH の初期ポーズは T ポーズであり，手前が正面向きとして鉛直上向きが y 軸方向，紙面奥が z 軸方向となる．なお，Hand Middle1 は中指の付け根の関節にあたる．一方，Pepper の初期ポーズは前ならえであり，手前が正面向きとして鉛直上向きが z 軸方向，紙面手前が x 軸方向となる．

BVH の肩関節の y 軸を回転させ，BVH と Pepper のポーズを前ならえに統一する．各関節の回転軸は BVH では 3 軸，Pepper では 2 軸で構成されるが，肩関節においては Pepper では考慮されない BVH の z 軸 (Pepper の x 軸) が及ぼす肩の動きへの影響が大きかったために，BVH の肩関節の z 軸である ArmZ を ShoulderRoll，ShoulderPitch に仮に表 1 に示すように加え，考慮した．表 1 に示す Hand は手の開閉を表す．Pepper への変換のために，今回は仮に中指の付け根の曲げ具合を表す Hand Middle1Z を用いた．また，Pepper の関節名の一部として表記される Yaw，Pitch，Roll はそれぞれ x，y，z 軸回転を表す．

学習データとなる人間の頭の向きは正面方向よりも下向きの場合が多く，Pepper に変換すると常に頭が下を向くことになる．また，腰の動きは Pepper に変換すると無理な姿勢を連続して行うことが多かった．これらの動きはジェスチャの評価に悪影響を及ぼすことが懸念されたために，Pepper への変換は両腕のみとした．

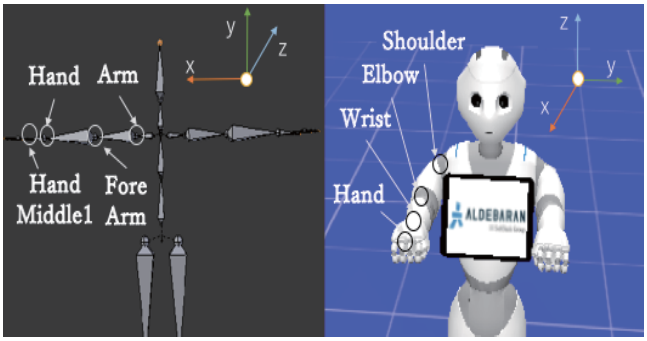


図 3　BVH と Pepper の座標軸と関節名

表 1　関節角度の対応表(符号上:右腕，符号下:左腕)

| Pepper の関節 | BVH の関節から Pepper の関節への対応式 |
|---|---|
| ShoulderRoll | ArmY ∓ 75 − ArmZ*5/6 |
| ShoulderPitch | − ArmX ∓ ArmZ*1/6 |
| ElbowRoll | ForeArmY (両腕同式) |
| ElbowYaw | ± ForeArmX |
| WristYaw | ± HandX |
| Hand | 90 + HandMiddle1Z (両腕同式) |

### 4.4 ジェスチャの自動生成に対する性能評価実験

#### (1) 実験設定

10 種類の音声データに対する original, predicted, mismatched の 3 種類のジェスチャの印象を検証する 1 要因 3 水準の実験を行った. original は音声データのジェスチャ, predicted は音声データから予測したジェスチャ, mismatched は別の音声データのジェスチャを音声の再生時間と一致するよう末尾を削除したジェスチャである.

参加者は 20 代理系大学生 28 名, 文系大学生 2 名の計 30 名 (男性 21 名, 女性 9 名) である. 音声データは **4.1** の (1) で述べられたデータセットのうち, モデルの学習に用いていない 10 秒ほどの音声を 10 種類選定した. 実験は Web 上で行い, 1 ページにつき動画を 1 つ表示する. 実験で使用した動画は, YouTube のサイト*3から閲覧できる. 3 条件の表示順は参加者 6 名毎にランダムに行い, 表示順による回答への影響を均一化した.

#### (2) 評価項目

ジェスチャの妥当性を「動きの自然さ」,「発話音声との時間的な整合性」,「発話音声との内容の整合性」の 3 項目から評価した. 以下に示すように, 項目毎に 3 つの小項目を設け, 合計 9 個の項目について 7 段階評価を行った.

・動きの自然さ
- Q1:ジェスチャは自然に見えた
- Q2:ジェスチャは滑らかであった
- Q3:ジェスチャに違和感はなかった

・発話音声との時間的な整合性
- Q4:ジェスチャのタイミングは発話音声と一致していた
- Q5:ジェスチャのスピードは音声発話と一致していた
- Q6:ジェスチャの緩急は発話音声と一致していた

・発話音声との内容の整合性
- Q7:ジェスチャは発話音声と一致していた
- Q8:ジェスチャは発話音声をよく表現していた
- Q9:ジェスチャは発話音声の理解を促進した

#### (3) 結果

ジェスチャの種類別の平均に対して, 一元分散分析を行ったところ, 主効果が認められた (Q1:F(2, 29)=4.766, p=.0121, Q2: F(2, 29)=20.27, p<.001, Q3:F(2, 29)=3.277, p=.0448, Q6:F(2, 29)=8.39, p=.001, Q7:F(2, 29)=17.18, p<.001, Q8:F(2, 29)= 14.96, p<.001, Q9:F(2, 29)=11.84, p<.001). Bonferroni の下位検定を行ったところ, Q2 では predicted が他条件に対して有意差が検出されるほど評価が高くなった. Q1, Q3 は original には平均評価値で下回るが, mismatched を上回る結果となった. Q5~Q9 においては predicted が他条件に対して有意に平均評価値が高くなる設問は一つも存在しなく, original に対して有意差が検出されるほど平均評価値が低くなる設問も存在した. 結果を図 4 に示す.

#### (4) 考察

- ジェスチャの自然さ

今回の結果から, 滑らかさにおいては十分な質のジェスチャが生成できていると考えられる. また, Q1~Q3 の結果から predicted は mismatched よりも自然なジェスチャであると考えられる.

- 時間的な整合性

今回の結果の要因として, 頻繁に手を動かしていたことが悪影響を及ぼしていると考えられる. また, ジェスチャの緩急においては, original に対して有意に平均評価値が低くなった. 原因として, 移動平均フィルタを用い動きが間引かれたことで, ジェスチャの速度が多少なりとも均一化されたことが考えられる.

*3 https://youtu.be/iINtbQUAhLI

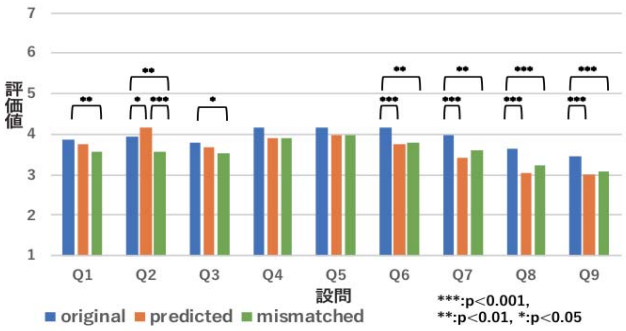


図 4 各設問の平均評価値

- 内容の整合性

全ての設問で, 平均評価値が他条件よりも下回る結果となった. このような結果の原因として, 提案手法により生成されたジェスチャは, 一定範囲で上下に手を動かすことが多く見受けられたことが挙げられる.

## 5. 講義代行システムの印象評価実験

### 5.1 実験設定

講義内容は芳尾ら [芳尾 2017] の論文から自動生成される. 芳尾らの論文では, 教科書と対応するスライドを基にシナリオを自動生成し, Pepper が講義代行するという内容が述べられている.

参加者は 20 代理系大学生 31 名 (男性:27 名, 女性:4 名) である. 実験後, アンケート用紙を配布し, 評価を行った. なお, ジェスチャの自動生成に対する性能評価実験時と同様に Pepper の動作は両腕のみとした.

あらかじめ録音した Pepper の音声を入力として, 4 章で述べられた手法を用い, ジェスチャを自動生成した.

### 5.2 評価方法

「講義全体」,「ジェスチャ」の二つの項目について評価した.

「講義全体」は GODSPEED [Bartneck 2008] を参考に「人間性」,「知性」,「好意」,「活発性」,「講義の有効性」という 5 つの測定概念を定めた. 表 2 に示すように, 概念毎に 3 組の形容詞対を設け, 全 15 組の形容詞対に対して SD 法 [末永 1987] を用いて 7 段階評価 を行った.

「ジェスチャ」は, 発話内容に対するジェスチャの妥当性を,「動きの自然さ」,「発話との時間的な整合性」,「発話と内容の整合性」を評価する 3 つの設問に対し, 7 段階評価を行った.

表 2 項目「講義全体」の概念毎の形容詞対

| 概念 | 形容詞対 |
|---|---|
| 人間性 | 機械的な-人間的な, 人工的-生物的, 偽物のような-自然な |
| 知性 | 愚かな-賢明な, 知的でない-知的な, 無知な-物知りな |
| 好意 | 嫌い-好き, 親しみにくい-親しみやすい, 不愉快な-愉快な |
| 活発性 | 熱心でない-熱心な, 非活発な-活発な, 活気のない-生き生きとした |
| 有効性 | わかりにくい-わかりやすい, 興味深くない-興味深い, 飽きのくる-飽きのこない |

### 5.3 結果

各項目の平均評価値の結果を表 3, 表 4, 図 5 に示す. 表 3 は概念に属する形容詞対の値の平均値となる.

講義全体では，高評価の概念は知性と活発性，低評価の概念は人間性と有効性，中央値 4 付近の概念の好意であった．ジェスチャでは，全ての設問で中央値 4 を下回った．

表 3　項目「講義全体」の概念毎の結果

| 概念 | 人間性 | 知性 | 好意 | 活発性 | 有効性 |
|---|---|---|---|---|---|
| 平均値 | 3.15 | 4.97 | 4.14 | 4.28 | 3.73 |

表 4　項目「ジェスチャ」の結果

| 設問 | 平均値 |
|---|---|
| ジェスチャは自然だったか | 3.19 |
| ジェスチャのタイミングは音声とあっていたか | 3.74 |
| ジェスチャと説明内容は一致していたか | 3.26 |

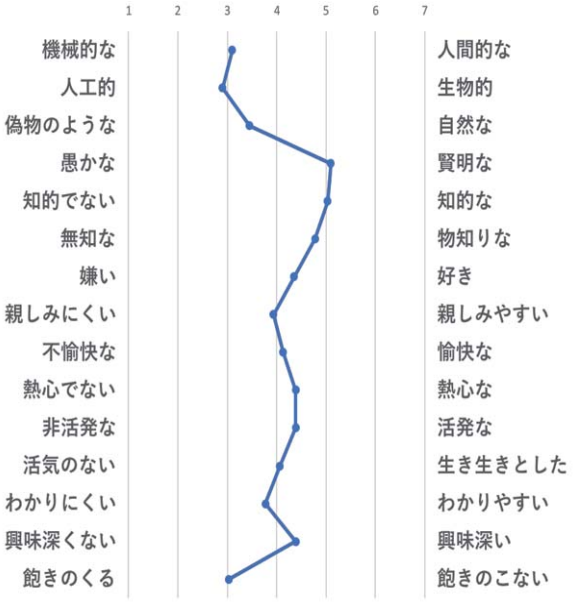


図 5 「講義全体」の 7 段階印象評価

## 5.4 考察

- 講義全体

点数の高かった概念である，活発性について考察を行う．活発性はジェスチャによる影響が大きく，常に手を動かしていたために活発な印象を与えたと考えられる．

次に，点数の低かった概念である人間性，有効性の考察を行う．これらの概念が低評価であった要因は，類似のモーションが多かった，話の内容とジェスチャの内容が合っていなかったといったことが挙げられる．

- ジェスチャ

3 つの設問のうち，「動きの自然さ」は 3.19 と最も低い値となった．低評価の要因は，まず，「ジェスチャは自然だったか」という内容が解釈の余地が大きい設問であり，他 2 つの設問を鑑みた「ジェスチャの総合的な自然さ」という評価も含まれていることが挙げられる．次に，動作が両腕のみであったことによる見た目の不自然さが及ぼした影響も低評価に繋がったと考えられる．また，生成されたジェスチャの速度に追いつけないために生じたガタつき等も低評価の原因の一つと考えられる．

## 6. まとめ

本稿では，自動生成によるジェスチャを用いた講義代行システムを提案した．

ジェスチャ自動生成に対する性能評価実験から，自動生成されたジェスチャは滑らかな動きであることが評価される一方，発話内容との時間的な整合性，内容の整合性は低評価であった．また，講義代行システムの印象評価実験では，活発な印象を与える結果となった．常に両腕を動かすことで，活発な身振りを表現できたと考えられる．一方，飽きやすい印象を与える結果となった要因としては似た動きが多いことが考えられる．

今後の課題として，発話内容の意味を取れるジェスチャを生成するために，ニューラルネットワークの改良とデータセットの増築を行いたい．データセットに関しては，openpose [Cao 2017] を用いて動画から回転角を導出し，データセットの増量を行うことを検討する予定である．また，自動生成する関節を両腕のみだけでなく全身の関節を対象に拡大することにより，見た目の違和感を除いていく予定である．

## 参考文献

[橋本 2012] Takuya Hashimoto, Igor M. Verner and Hiroshi Kobayashi.: “Human-Like Robot as Teacher’s Representative in a Science Lesson: An Elementary School Experiment.” Robot Intelligence Technology and Applications, pp.775–786, 2012.

[橋本 2006] Takuya Hashimoto, Sachio Hitramatsu, Toshiaki Tsuji and Hiroshi Kobayashi.: “Development of the Face Robot SAYA for Rich Facial Expressions.” In international joint conference on artificial intelligence, pp.5423–5428, 2006.

[Xu 2014] Junchao Xu, Joost Broekens, Koen Hindriks and Mark A. Neerincx.: “Effects of Bodily Mood Expression of a Robotic Teacher on Students.” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.2614–2620, 2014.

[日和 2017] 日和航大, 荒木健治, 長谷川大: “ヒューマノイドロボットを用いた講義代行システムにおけるジェスチャの考察”, 電気・情報関係学会北海道支部連合大会, pp.220‐221, 2017.

[竹内 2017] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta and Kazuhiko Sumi.: “Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-Directional LSTM.” In Human-Agent interaction (HAI) symposium, pp.365–369, 2017.

[竹内 2017] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa and Hiroshi Sakuta.: “Creating a gesture speech dataset for speech-based automatic gesture generation.” In HCI International 2017–Posters’ Extended Abstracts: 19th International Conference, HCI International 2017, pp.198–202, 2017.

[芳尾 2017] 芳尾哲, 荒木健治.: “ヒューマノイドロボットを用いた講義代行システムのための講義シナリオ文自動生成手法”, 電気・情報関係学会北海道支部連合大会, pp.218‐219, 2017.

[Bartneck 2008] Christoph Bartneck, Dana Kulic and Elizabeth Croft.: “Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots.” In Proceedings of the Metrics for Human-Robot Interaction Workshop in Affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp.71–81, 2008.

[末永 1987] 末永俊郎. “社会心理学研究入門”, 東京大学出版会, 1987.

[Cao 2017] Zhe Cao, Tomas Simon, Shin-En Wei and Yaser Sheikh.: “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.” In IEEE, pp.1302–1310, 2017