CJ-005

# 対象物のイメージに基づく図像的ジェスチャの形状推定
# Estimating Iconic Gesture Forms based on Entity Image Representation

二瓶 芙巳雄† 中野 有紀子† 東中 竜一郎‡ 石井 亮‡
Fumio Nihei Yukiko I. Nakano Ryuichiro Higashinaka Ryo Ishii

## 1. はじめに

対面コミュニケーションにおけるハンドジェスチャはスピーチに結びつく，あるいは付随するものであり，非言語コミュニケーションシグナルとしての基本的なものである．バーチャルエージェントやコミュニケーションロボットのジェスチャ生成の研究 [1]においてもこの点を重視し，発話内容の理解を助けるジェスチャを選択することが重要であると考えられる．

コミュニケーション研究において様々なジェスチャの種類が定義 [2]されているが，McNeil の分類では，発話の中で言及された具体的な対象物のイメージを提示する図像的ジェスチャ (iconic gesture) や，抽象的な概念を伝える隠喩的ジェスチャ (metaphoric gesture) などのジェスチャ種類が定義されている [3]．バーチャルエージェントやコミュニケーションロボットがこれらのジェスチャを表出することの効果として，図像的ジェスチャを適切に表出することにより，ユーザの会話内容の理解度の向上や [4]，教育場面での記憶力の向上に役立つこと [5] が知られている．従って，エージェントやロボットの図像的ジェスチャの自動生成は重要な研究課題である．

本研究では，ジェスチャを通じて，対象物の形や大きさといった付随的な情報をより詳しく伝達できるバーチャルエージェントの実現をめざし，対象物に対する図像的ジェスチャの形態の決定に焦点を当てる．ジェスチャ生成に関する従来研究として，仮想空間での会話を分析し，仮想空間内の対象物の幾何学的な情報から図像的ジェスチャの形態を決定する手法が提案されている [6, 7]．この研究では，対象物の幾何学的・空間的特徴を表現するためにイメージ記述 (image description) の考え方を採用している．同様の手法として，Ravenet ら [11]は Image Schema を提案し，ジェスチャの特徴に Image Schema を対応付けた辞書を定義し，隠喩的ジェスチャを生成している．しかし以上の研究では，手動で割り当てられた特徴，あるいは手動で定義された辞書を使用している．

本研究では，多様な図像的ジェスチャを自動生成するための新たな試みとして，ディープニューラルネットワーク (DNN) を用いて，ある対象物の概念表現 (image representation) を画像から自動で獲得する．さらに，この概念表現を用いてハンドジェスチャの形態決定を学習することにより，ジェスチャ辞書を自動で作成する技術を提案する．具体的には，ある対象物の画像の集合から典型的な概念表現を生成し，これに基づき，対象物についてアノテータがイメージするハンドジェスチャの形状を 7 種類の中から一つを決定する DNN を提案する（図 1）．また，モデルにより決定されたジェスチャとその対象となる単語のマッピングを定義したジェスチャ辞書も作成する．様々な対象物に対してジェスチャの形態が保有された大規模辞書が構築されれば，バーチャルエージェントがジェスチャを通じて対象物の形や大きさといった付随的な情報をより詳しく伝達でき，実用的な方式として利用できる可能がある．

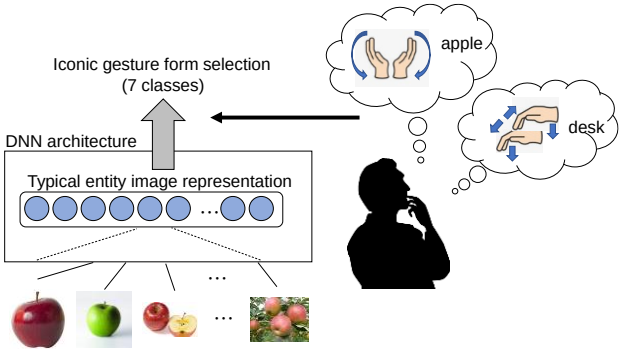


図 1　DNN による概念表現獲得手法.

## 2. データ

日本語シソーラス[1]から具体物か無生物に分類される 18,580 単語を対象物として抽出し，それらのうち 5%からモデルを作成する．学習データを作成するために，選定された対象物に対して最適なジェスチャ形状を，アノテータがラベル付けした．

その後，各単語について複数の画像を収集することで，対象物の単語，画像集合，そして単語に対する最適なジェスチャ形状の組みを作成した．

### 2.1 ジェスチャ形状アノテーション

ジェスチャの形状を次の基本的な形状 7 種類として分類した：正方形，垂直四角形，水平四角形，円，垂直楕円，水平楕円，線状．5 名のアノテータは，ある対象物の形状を手でジェスチャするときの最適な形状を 7 種類の中から 1 つ選ぶように指示された．ここでアノテータは，対象物を写した画像を見ずに作業した．このように単語だけを見ることは，アノテータ自身の頭の中にある対象物についての典型的なイメージや概念に基づいてジェスチャの形状を選ぶことを意味する．もしも人々がある対象物に対して似たようなイメージ（塔は縦長である，丸い机もあるかもしれないが机は典型的には横長の長方形である）を共有していれば，その形の判断は一致するはずである．結果として，97%の対象物において，5 人のアノテータのうち 2 人のアノテーションは一致していたが，以降の分析では，過半数

† 成蹊大学理工学部　Faculty of Science and Technology, Seikei University
‡ 日本電信電話株式会社　NTT メディアインテリジェンス研究所　NTT Media Intelligence Laboratories, NTT Corporation

[1] 日本語語彙大系: http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTaikei/index.html

表 1　エンティティ数と画像数.

| ジェスチャ形状 | エンティティ数 | 画像数 |
|---|---|---|
| 四角 | 118 | 7,509 |
| 垂直四角 | 112 | 8,379 |
| 水平四角 | 91 | 5,048 |
| 円形 | 154 | 9,520 |
| 垂直円形 | 30 | 1,702 |
| 水平円形 | 36 | 1,687 |
| 線形 | 56 | 3,486 |

のアノテータにおいて対象物の形状のイメージが共有されていたことを意味する，5 人中 3 人のアノテータの判断が一致した 597 の対象物を使用した.

### 2.2 画像収集

597 対象物の画像をそれぞれ最大 200 枚ずつ，Google 画像検索[2]により収集した．200 枚の画像は検索エンジンが上位の検索結果として表示したものである．検索キーワードには対象物の単語を使用した．しかし多くの単語は複数の意味を持つ．例えば，検索キーワードが「トランプ」の場合，「ドナルド・トランプ」や「トランプカード」の画像が収集された．そのため，各対象物の国語辞典における定義し，定義に一致しない画像を削除した．さらにオブジェクトの上に文字が重なっているなど，対象物の外観以外の不要な情報を含む画像も削除した．その結果 597 対象物について 37,331 枚の画像が得られた．表 1 にはジェスチャ形状カテゴリごとの対象物の種類数と対象物の画像数を示す.

## 3. ジェスチャ形状決定手法

ジェスチャ形状の決定を 2 つのタスクに分け，それぞれのタスクに対応した推定モデルを作成した．一つ目のモデルである Basic Form モデル（以下 BF モデル）は，画像の集合から四角形，円形，線形の 3 つの基本的なジェスチャ形状の中から 1 つを決定する 3 クラス分類モデルである．四角形と円形については，二つ目のモデルである Vertical-Horizontal モデル（以下 VH モデル）が，BF モデルと同様に画像の集合からその形状が等辺形，縦長，横長のいずれかを決定する．これも 3 クラス分類問題である.

ネットワークのアーキテクチャを図 2 に示す．このネットワークは BF モデルと VH モデルの両方で使用されている．BF モデルは，画像の集合$\mathrm{I} = (\mathrm{I}_1, \mathrm{I}_2, \ldots, \mathrm{I}_n)$から基本ジェスチャ形状$\mathrm{Y_{BF}}$を推定する．同様に，VH モデルは，画像集合 I から等辺形/縦長/横長を示す$\mathrm{Y_{VH}}$を決定する．入力画像サイズは 224×224×3（幅 224×高さ 224 の RGB 画像）である．画像のエンコーダとして事前学習済みの VGG-16 モデル[8]を使用し，各入力画像から$\mathrm{m} = 1 \ldots \mathrm{n}$の 4096 次元画像特徴量$\mathrm{F_m}$を抽出した．対象物の典型的な概念表現を作成するために，画像特徴量の集合から平均ベクトルを計算する.

$$F_m = VGG(I_m)$$
$$F = Average(F_1, F_2, \ldots, F_n)$$

ここで n は，対象物の概念表現を作成する際に使用される画像の数である．n = 1 の場合，モデルは特定の画像に基づいてジェスチャ形状を決定する.

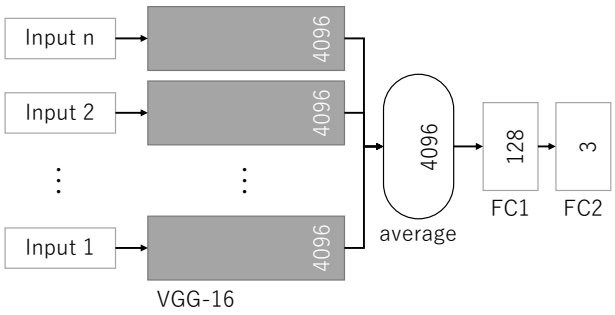


図 2　ネットワークアーキテクチャ.

対象物の概念表現 F は全結合層 FC と FC1 に供給され，活性化関数 ReLU を通して 128 次元ベクトルに射影される．続いて FC2 では，BF モデルからの予測結果$\mathrm{Y_{BF}}$と VH モデルからの予測結果$\mathrm{Y_{VH}}$の両方が softmax 関数を介して出力される．softmax 出力に対してカテゴリカル交差エントロピーを損失として算出し，確率的勾配降下法（SGD）により FC 層を訓練する.

## 4. 実験

3 章で収集したジェスチャ形状ラベルが与えられた画像データを使用しモデルを訓練・評価する．ネットワークに供給する画像の枚数に伴う推定性能の変化を明らかにすることで，最適な入力画像数を調査する.

モデルの訓練に使用したデータは 328 対象物についての 20,721 画像，検証には 149 対象物についての 9,597 画像，テストには 120 対象物についての 7,013 画像である．データセットに同一の画像が 2 回以上含まれている可能性は低い.

対象物の概念表現を作成する際に，入力画像の数を 2, 4, 6, 8, 10, 12 に変更することで，6 つのモデルを作成した．これらのモデルは 4 章で説明したように，VGG-16 が出力する特徴ベクトルの平均値を取ることで対象物の概念表現を計算する．ベースラインとしてランダムに選ばれた 1 枚の画像をネットワークに入力し，その画像から対象物の概念表現を作成するモデルを設定した.

収集できた画像数は対象物によって異なる．すなわち訓練インスタンスを単純に作成する場合，多く画像を収集できた対象物に強く影響されたモデルが訓練されてしまうことが予想される．そこで収集できた画像数に伴うモデルへの影響を制御するため，次の 3 段階の手続きで訓練インスタンスを作成した；1) 対象物ごとの画像集合から 64 枚の画像をランダムに選択する，2) 64 枚の画像から n 枚の画像を選択し 1 インスタンスを生成する，3) 手続き 2 を 63 回繰り返す，すなわち対象物についてのインスタンスを計 64 作成する．ここでベースラインモデル（n = 1）の場合，各インスタンス（画像）は一意である．一方 $\mathrm{n} = 2, 4, 6, 8, 10, 12$の場合，同じ画像が異なるインスタンスに含まれる．また検証用のインスタンスは，対象物ごとの分類性能を検証するため，上述の 3 段階の手続きの内 3 段階目を除外し作成された.

[2] https://github.com/hardikvasa/google-images-download

表 2　BF, VH モデルの Accuracy.

| 画像数 | BF モデル | VH モデル |
|---|---|---|
| 1 (ベースライン) | 0.742 | 0.627 |
| 2 | 0.782 | 0.654 |
| 4 | 0.804 | 0.709 |
| 6 | 0.816 | 0.728 |
| 8 | 0.803 | 0.738 |
| 10 | 0.816 | 0.752 |
| 12 | 0.806 | 0.764 |

表 3 二段階推定による推定性能.

| モデル | Accuracy |
|---|---|
| BF-1→VH-1 (ベースライン) | 0.495 |
| BF-2→VH-2 | 0.532 |
| BF-4→VH-4 | 0.577 |
| BF-6→VH-6 | 0.608 |
| BF-8→VH-8 | 0.598 |
| BF-10→VH-10 | 0.622 |
| BF-12→VH-12 | 0.629 |

### 4.1 モデル性能

テストデータにおけるモデルの性能を表 2 に示す．テストデータのインスタンスは，検証用インスタンスの作成と同一の方法で作成された．

#### 4.1.1 BF モデル

表 2 の 1 列目に，四角形，円形，線形の 3 つのジェスチャを出力する BF モデルの性能を示す．ベースラインモデル BF-1（入力画像数 1）に加えて，次の 6 つのモデルを作成した：入力画像の枚数がそれぞれ 2, 4, 6, 8, 10, 12 である BF-2, BF-4, BF-6, BF-8, BF-10, BF-12．表に示すように，6 つのモデルはすべてベースラインモデル（0.742）を上回った．このことは，単にランダムに選択された 1 枚の画像に基づいてジェスチャ形状を決定するよりも，複数の画像が必要であることを示している．一方，最も性能の良いモデルは BF-6 と BF-10 であり，いずれの Accuracy も 0.816 であった．性能には若干のばらつきがあるものの，入力画像数が 4 枚以上の場合には 80%以上の性能が得られることが明らかになった．以上の結果から，最低限 4 枚以上の画像を用いて対象物の概念表現を作成すると，人々が対象物を表現する際に選択する基本的なジェスチャの形をより正確に推定できることが示唆された．

#### 4.1.2 VH モデル

表 2 の 2 列目は，BF モデルにおいて四角形と円形のジェスチャ形状が選択された場合に，ジェスチャ形状が縦長，横長，等辺形のいずれであるかを判定する VH モデルの性能を示したものである．VH モデルの評価は，BF モデルと同様に行った．我々は入力画像数を変え，次の 6 つのモデルを作成した：VH-2, VH-4, VH-6, VH-8, VH-10, VH-12．表に示すように，入力画像数が多いほど Accuracy が高いことがわかる．最も性能の低いモデルはベースラインモデル（0.627），最も性能の高いモデルは VH-12（0.764）であった．入力画像が 4 枚以上の場合には，70%以上の性能が得られた．したがって，BF モデルと同様，最低限 4 枚以上の画像を用いて作成された概念表現を用いたモデルでは，ジェスチャの縦横比の特性を決定する際に優れた性能を発揮していることを示している．

#### 4.1.3 二段階ジェスチャ形状選択

提案手法の総合評価として，ジェスチャ形状を 7 形状に分類する際の有効性を検討した．まずテストデータを BF モデルに供給し，基本形状（円形，四角形，線形）を決定する．基本形状として円形や四角形が選択された場合は，その後 VH モデルを用いて，縦長，横長，等辺形のいずれかを決定する．この 2 段階の処理を行った場合の性能を表 3 に示す．例として，BF-8→VH-8 のセルは，第 1 ステップで BF-8 モデルを使用した後，第 2 ステップで VH-8 モデルを使用した場合の Accuracy を示している．表に示すように，すべてのモデルがベースラインモデルを上回る性能を示した．さらに BF-10→VH-10 及び BF-12→VH-12 では，62%以上の Accuracy が得られている．この結果は，7 種類のジェスチャを使い分けようとすると，数枚の写真だけでは選択性能が不十分であるが，10 枚程度の画像を入力とすると性能が大きく向上することを示唆している．

比較として，Kadono らは，円形，四角形，線形，その他の 4 種類のジェスチャを分類する手作業で定義した特徴量に基づくモデルを提案している [9]．円形，四角形，線形のモデルの平均 F 値は 0.55 であった．本研究では異なるデータセットを用いたためモデルの性能を直接比較することはできないが，本研究で用いた BF-10 モデルの 3 つの基本形の平均 F 値は 0.77 であり，文献 [9]よりも良い性能が得られた．

## 5. 提案モデルの ECA システムへの組み込み

提案手法の応用として，ジェスチャ辞書を作成し，Embodied Conversational Agent (ECA) に統合した．提案手法を対象物の画像の集合に適用し，対象物に対してジェスチャ形状を割り当てることで辞書を作成した．辞書自体は静的なデータベースであり，事前に作成する．

ECA システムには，動作生成モデル [10]が含まれている．このモデルはテキストを入力として，品詞タグや対話行為などの自然言語解析特徴と，シソーラスから得られる意味的関係や語彙的関係を計算する．そして頭部動作，顔の表情，手の動作を含む 8 種類の動作を決定する．ハンドジェスチャについては，ジェスチャを実行する際に，図像的，隠喩的，ビートなどのジェスチャカテゴリを選択する．動作生成システムが特定の単語で図像的ジェスチャを生成することを決定すると，システムがジェスチャ辞書を参照してジェスチャ形式を選択する．選択されたエージェントの動作は，動作スケジュール情報に基づいて，合成された音声とともにアニメーションエンジンを介してレンダリング

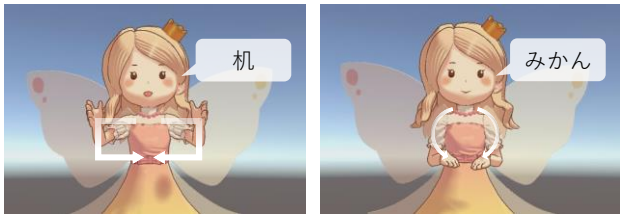


図 3　図像的ジェスチャ生成のスナップショット.

される．

このように ECA システムでは，テキストを入力として，音声と同期したエージェントの動作アニメーションを自動的に生成する．例として，「机の上にみかんがある」という文がシステムに入力されると，システムは「みかん」と「机」に図像的なジェスチャを生成することを決定する．そして，ジェスチャ辞書を参照して，「みかん」には円形のジェスチャを，「机」には横長の長方形のジェスチャを選択する．図 3 はこれら 2 つのジェスチャに対するエージェントの動作のスナップショットを示す．

## 6. まとめと今後の課題

図像的ジェスチャの主要な機能は，参照元の形状または物理的特徴を描写することである．さらにジェスチャの形状は，対象物の意味や概念など，ある種の概念表現と深く関係していることが知られている．しかし，ジェスチャ生成のための理想的な表現は未だに不明であり，特にジェスチャの自動生成（ジェスチャの形状決定）は困難な課題となっている．本研究では，対象物の物理的な外観に着目して，この問題に取り組んだ．

我々は DNN 技術を用いて，複数の画像から対象物の概念表現を生成した．実験では 10〜12 枚の画像から計算された概念表現がジェスチャ形状の決定に対して最も良好な結果を示し，1 枚あるいは数枚の画像から計算された概念表現はジェスチャ形状の決定には不十分な結果となった．また重要な発見として，10 枚程度の画像から作成された概念表現はジェスチャ形状の予測性能を大幅に向上させることを明らかにしたことが挙げられる．このことは提案手法の実用可能性を示唆している．

さらに提案手法を用いてジェスチャ辞書を作成し，エージェントが図像的ジェスチャを行うアニメーションを生成することで，提案手法の適用可能性を例示した．本研究では 3 章で述べたように，辞書の定義に合わない画像や，文字がオブジェクトに重なっている画像を手作業で削除しており，ジェスチャ辞書の開発において以上の問題が発生しないことを前提にしている．ここで画像上の文字は文字認識技術を用いて自動的に削除することが可能である．また対象物についての画像を収集する際には，開発者に対しては単語の定義を提供することが望ましい．

今後は，実装したエージェントシステムのユーザスタディを実施し，生成されたジェスチャがユーザにとって理解しやすく，エージェントとのコミュニケーションに役立つかどうかを検証する．また，「表現」の概念を拡大する必要がある．現在の方法では画像だけから表現を計算するが，概念は画像だけに由来するものではない．Image Schema [11]や IDTs representation [6]で議論されたように，ECA システムのジェスチャ形式決定モデルにおいて，言語情報や文脈情報を組み込む必要がある．

### 謝辞

本研究の一部は JSPS 科研費 JP19H01120，JP19H04159 の助成を受けたものです．

### 参考文献

[1] Cassell, J. Bickmore, T. Campbell, L. Vilhjálmsson, H. and Yan, H. 2000. Embodied Conversational Agents. . MIT Press, Cambridge, MA, USA, 29–63. Retrieved from http://dl.acm.org/citation.cfm?id=371552.371555

[2] Knapp, ML. Hall, JA. and Horgan, TG. 2013. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing.

[3] McNeill, D. 2006. Gesture and Communication. *Encyclopedia of Language & Linguistics*, 1910: 58–66. DOI: http://dx.doi.org/10.1016/b0-08-044854-2/00798-7

[4] van Dijk, ET. Torta, E. and Cuijpers, RH. 2013. Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics* 5, 4: 491–501.

[5] Bergmann, K and Macedonia, M. 2013. A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance. *Proceedings of the 13th International Conference on Intelligent virtual agents (IVA 2013)*, 139–148.

[6] Bergmann, K and Kopp, S. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 361–368.

[7] Kopp, S. Tepper, PA. Ferriman, K. Striegnitz, K. and Cassell, J. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. In *Conversational Informatics*, T. Nishida (ed.). John Wiley, 133–160.

[8] Simonyan, K and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*.

[9] Kadono, Y. Takase, Y. and Nakano, YI. 2016. Generating iconic gestures based on graphic data analysis and clustering. *ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 447–448. DOI: http://dx.doi.org/10.1109/HRI.2016.7451799

[10] Ishii, R. Katayama, T. Higashinaka, R. and Tomita, J. 2018. Generating Body Motions using Spoken Language in Dialogue. *Proceedings of the 18th International Conference on Intelligent virtual agents (IVA 2018)*, 87–92. DOI: http://dx.doi.org/10.1145/3267851.3267866

[11] Ravenet, B. Pelachaud, C. Clavel, C. and Marsella, S. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in Psychology* 9, JUL. DOI: http://dx.doi.org/10.3389/fpsyg.2018.01144