2D4-2

画像データに基づく図像的ジェスチャの自動生成

Automatic Generation of Iconic Gestures based on Image Data

門野 友城[*1] 高瀬 裕[*2] 中野 有紀子[*2]
Yuki Kadono Yutaka Takase Yukiko Nakano

[*1] 成蹊大学理工学研究科
Graduate School of Science and Technology, Seikei University

[*2] 成蹊大学理工学部
Faculty of Science and Technology, Seikei University

In human-agent interaction, nonverbal behavior has an important role such as improving the comprehensibility of conversational content. However, gesture generation is one of the most difficult tasks in humanoid interfaces. This study proposes a method of generating iconic drawing gestures using image processing and machine learning techniques. We collected a set of graphic images for over 1000 objects and classified the objects into 4 types of shapes. By implementing a gesture shape decision mechanism, we also built a system that takes a sentence as the system input and produces hand gesture animations that are synchronized with synthetic speech.

1. はじめに

ヒューマノイドインタフェースは人のように，体を使用した表現ができる．体を使用した表現の一つとしてハンドジェスチャが存在する．対面対話でのインタラクションだけではなく，アニメーションエージェントやヒューマノイドロボットによるハンドジェスチャは会話内容の理解の向上に役立つことが知られている[Rogers 1978] [Dijk et al. 2013]．このことより，ジェスチャ生成はヒューマノイドインタフェースの構築をする上で重要である．特に美術館などでガイドをしたり，会話で道案内をしたりする場合などにおいて，ジェスチャ表現は必要不可欠である．しかし，ジェスチャの自動生成はとても難しい問題となっている．方法の一つとして，ジェスチャの形を登録して，各単語にジェスチャの形を割り当てるような，ジェスチャの辞書を手作業で定義するというものがある．しかし，各単語に手作業でジェスチャの形を割り当てるには膨大な時間とコストがかかり，作業に限界がある．

本研究では，画像処理とクラスタリング手法を適用することで，[McNeill 1992]が提案した物の形をジェスチャによって表現する「図像的ジェスチャ」の自動生成の方法を提案する．また，入力として文字列を受け取り，発話に共起したハンドジェスチャアニメーションを出力するバーチャルエージェントシステムの構築も行う．

2. 関連研究

2.1 図像的ジェスチャの分類

[Lücking et al. 2013]は図像的ジェスチャと指示的ジェスチャに焦点を当てて研究を進めている．この研究で，実験参加者の発話とジェスチャの注釈付けを行った SaGA コーパスを調査した．その中で彼らは図像的ジェスチャ及び指示的ジェスチャを以下の 8 種類に分類した．

- Indexing : ジェスチャスペースで位置を指さす
- Placing : 物体をジェスチャスペースに置いたり並べたりする動作
- Shaping :空中に形の輪郭を描く
- Drawing : 形の概略をたどる
- Posturing : 手で物体のモデルや代わりを示す
- Sizing : 距離や大きさを示す
- Counting : 指で数を数える
- Hedging : 不確実な指摘(肩をすくめるなど)

今回は Drawing ジェスチャに焦点を当て，様々な物体の形状をたどるようなジェスチャの自動生成を目指した．

2.2 ジェスチャ生成の手法

ジェスチャ生成の研究として[Sadeghipour et al.2014]は実験で被験者より行われた図像的ジェスチャの構造を調査することで，図像的ジェスチャの組成のパターンを抽出した．その中で，Feature-based Stochastic Context-Free Grammars (FSCFG)を提案した．FSCFG はジェスチャ生成のルールとして使用される．しかし，本研究では人のデータを集めず，画像処理を利用することにより Drawing ジェスチャで描くための物体の形の特徴を抽出する．例えば，キーボードの画像からその形状が「四角」であることが把握できれば，それに応じたジェスチャをエージェントに行わせることが可能となる．

3. ジェスチャの形の決定

図像的ジェスチャを行場合，細かく物体の形を描くことはなく，大まかに形の特徴を描くことが多い．この仮定に基づき，本研究ではジェスチャ対象となる様々な物（オブジェクト）の外形を 4 種類に分類する手法を提案する．これによってエージェントにジェスチャを行わせるために必要なモーション作成に要する時間的コストが削減できる．分類する形状とそれに応じたジェスチャは以下の通りである．

- Circle ： 丸い形のジェスチャをする
- Rectangle ： 四角い形のジェスチャをする
- Linear ： 線の形のジェスチャをする
- Other ： 上記の何れにも当てはまらない場合で，ジェスチャは行わない

本研究では単純な画像処理に加えて，クラスタリング・機械学習を利用し，物体を上記 4 つの形状へと分類した．図 1 に提案手法の流れを示す．本手法は大きく 4 つの段階に分けられる．以下に詳細を示す．

連絡先：門野 友城，成蹊大学大学院 理工学研究科，
dm156203@cc.seikei.ac.jp

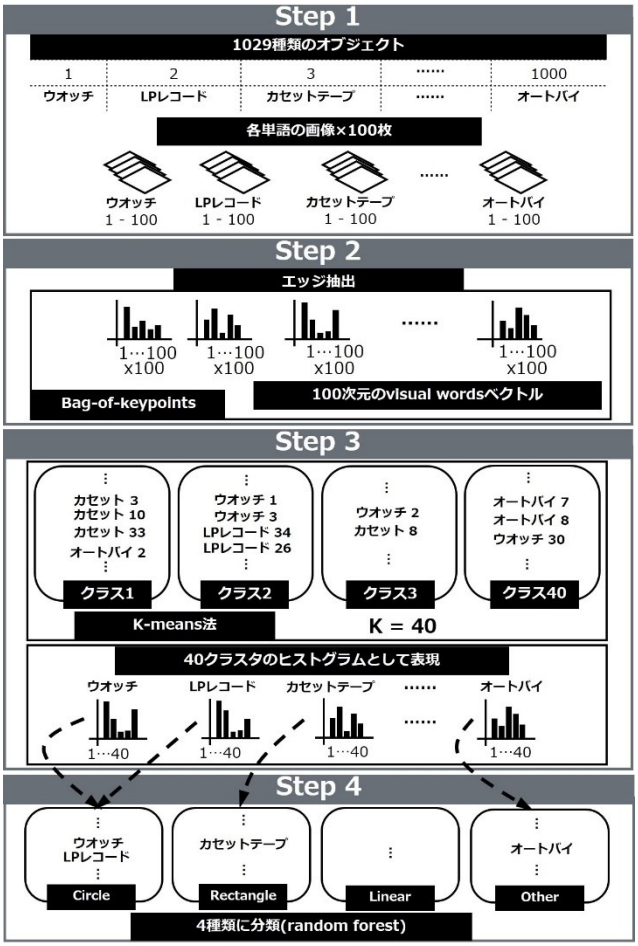

図 1. 形決定の流れ

Step1：最初に日本語語彙大系から 1029 のジェスチャとして表現できる可能性のある名詞(乗り物・家具などの人工物)をランダムに選択した. ここで利用する日本語語彙大系は約 30 万の日本語を収録した NTT コミュニケーション科学基礎研究所監修の大規模日本語辞書である. そして, 各単語について約 100 枚ずつの画像を Microsoft Bing Search により提供されている画像検索 API を利用して, インターネット上から収集した.

Step2：次に, Step1で収集した各画像から物体のエッジを抽出した. エッジを抽出した画像に OpenCV[*1] で実装されている [Lowe 2004]によって考案された SIFT(Scale Invariant Feature Transformation)アルゴリズムを使用することで, keypoint と呼ばれる画像の特徴点を抽出した. 本来, SIFT アルゴリズムではエッジ上の点は特徴点として適切ではないが, エッジ画像に対して SIFT を適用することでその外形の分類に利用できると考えた. その後, 各画像から特徴量のスケールの大きいもから 100 個ずつ keypoint を選択し, これら 102900 個の SIFT 特徴量を k-means 法 (k=100)を利用しクラスタリングすることで Code book と呼ばれる画像の辞書を作成した. この手法は bag-of-keypoints と呼ばれる[Csurka et al. 2004]. この Code book を利用すれば, 各画像を 100 次元のベクトルとして表すことができる. 具体的には各画像における特徴点それぞれが, 上記 100 次元クラスタのうち, どのクラスタに属するものかを求め, 100 次元のヒストグラムを作成する. 各クラスタは visual word と呼ばれるクラスタ重心ベクトルで代表され, 特徴点と各 visual word とのユークリッド距離を算出し, 最近傍クラスタをその特徴点が属するクラスタとし, その頻度を数える.

Step3：本研究の目的は与えられた画像の形を判定することではなく, 各オブジェクトの形を決定するために分類することである. そこで, 約 100 枚の画像から成る各オブジェクトの特徴を表現する必要がある. この目的を達成するために, k-means 法 (k=40)を Step2 で生成した, 100 次元ベクトルで表現された各画像に適用した. これによって 40 次元のヒストグラムとして各オブジェクトを表現することができる. ここで私たちは分布の似たクラスタのオブジェクトは同じ典型的な形を持つと仮定した.

Step4：最後に機械学習を利用し, 各オブジェクトを先述の 4 つの形状に分類した. 具体的には, 1029 のオブジェクトからランダムに 200 のオブジェクトを選択し, 教師データとして本章の最初に決定した基本となる 4 種類の形を割り当てた. 特徴量には Step3 で生成した, オブジェクト毎の 40 次元ベクトルを使用し, Random Forest アルゴリズムを適用することによって 4 値分類モデルを作成した. また, 教師データとは別に 100 のオブジェクトをテストデータとし, 分類精度の評価を行った. その結果を表 1 に示す. チャンスレベルのベースラインを 0.25 とすると, ここでの結果はモデルの分類は 4 つのクラスに対して良い結果を示した.

表 1. 形状分類のためのモデルの評価

クラス	Recall	Precision
Circle	0.56	0.50
Rectangle	0.52	0.531
Linear	0.56	0.667
Other	0.6	0.608

4. ジェスチャ生成

最後に, バーチャルエージェントシステムに提案したジェスチャ自動生成メカニズムを実装した. 各オブジェクトの単語へジェスチャ形状と大きさを定義したジェスチャ辞書を作成した. ジェスチャ形状は第 3 章で提案した方法を使用することで自動的に割り当てられる. 大きさは人手で 1 から 5 のパラメータを割り当てた. 例として, バスケットボールには 3(中間の大きさ), スポンジボールへは 1 というパラメータ (最小) を割り当てた. システム統合のためのアニメーションエンジンとして, Unity*2 を使用した. Unity は多くのプラットフォーム上で動作するゲームエンジンである. ジェスチャ生成のプロセスを以下に述べる. また, システム構成図を図 2 に示した.

(1)ユーザによって任意の文字列が入力されると, 形態素解析により名詞を抽出する. 各名詞をシステムがジェスチャ辞書に存在するか検索する. もしジェスチャ辞書に名詞が合った場合, そのパラメータに応じて, ジェスチャの形状と大きさを決定する. そのとき, ここで得た情報を元に最初に入力された文字列と共にジェスチャすべき名詞にタグを付与した XML ファイルをエージェントモジュールへ送信する.

(2)エージェントモジュールは Microsoft Speech API (SAPI) を使用した text-to-speech(TTS) 機能を持ち, 同時にリップシンクアニメーションも生成することが可能である. このモジュールは上記で特定したジェスチャの形と大きさによってエージェントのジェスチャアニメーションを選択する. また, このモジュールではジェスチャと, リップシンクのための visemes ID を含んだアニメーションのタイムスケジュールも決定する. (1)で生成された XML ファイルを元にジェスチャタグ付きのテキストを SAPI に送信すると, 結果として visemes ID とジェスチャ ID を得る. SAPI で発話を行うと共に visemes コマンドとジェスチャコマンドを

*1 http://opencv.org

*2 https://unity3d.com/jp

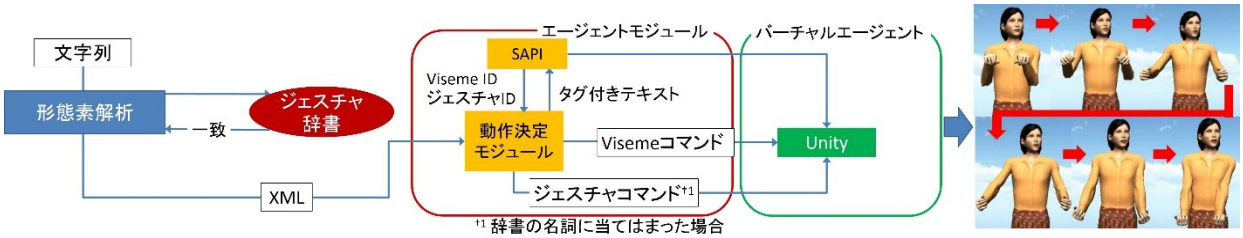

図 2. システム構成図

Unity に送信する. これにより, このシステムはバーチャルエージェントにリップシンクを伴った発話と適切なタイミングで適切なジェスチャをすることを実現する. アニメーションは全て Unity 上で行われる. 図 2 の右のスナップショットは実際にエージェントが丸いものを表現する circle ジェスチャを行っている場面である.

5. おわりに

この研究により画像処理とクラスタリング手法を利用した図像的ジェスチャ生成の手法を提案した. また, バーチャルエージェントシステムへこの手法を実装することにより, 発話と同時にオブジェクトを表すジェスチャを生成することを実現した.

今後の課題として, ジェスチャの種類を増やすことが必要だと考える. 加えて, 今回使用しているジェスチャが本当に適切であるか, 例えば CD などの丸いオブジェクトを表現する際に両手を使用する, などを検討する必要もある.

また, 実際にジェスチャを付与したバーチャルエージェントを利用することでユーザが発話内容の理解の向上に繋がるのかを調査するために評価実験を行う必要もあると考える.

将来的に, ショッピングモールでの案内や道案内などのバーチャルエージェントに適用することで, 会話の手助けになればと考えている.

謝辞

本研究は, 科学技術振興機構 (JST) 戦略的想像研究推進事 業 (CREST)「実践知能アプリケーション構築フレームワーク PRINTEPS の開発と社会実践」の支援によって実施した.

参考文献

[Rogers 1978] Rogers, W.: The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior within Utterances, Human Communication Research, 5: pp. 54-62, 1978.

[Dijk et al. 2013] Dijk, E. T., Torta, E., Cuijpers, R. H.: Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction, International Journal of Social Robotics, 5(4):491–501, Sept. 2013.

[McNeill 1992] McNeill, D: Hand and mind: what gestures reveal about thought, 1992.

[Lücking et al. 2013] Lücking, A. Bergman, k., Hahn, F., Kopp, S. and Rieser, H.: Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications,, Journal of Multimodal User Interfaces 7(1-2): 5-18, Springer ,2013.

[Sadeghipour et al.2014] Sadeghipour, A. , Kopp, S.: Learning a Motor Grammar of Iconic Gestures, Proceedings of the 35th Annual Meeting of the Cognitive Science Society, Cognitive Science Society,2014.

[Lowe 2004]Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, Vol.60, No.2, pp.91–110,2004.

[Csurka et al. 2004] Csurka, G., Bray, C., Dance, C. and Fan, L. Visual categorization with bags of keypoints, Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, pp. 1–22, 2004.