1E1-02

台本に基づく会話エージェントのジェスチャ自動生成

Automatic Gesture Generation for Conversational Agent Based on Scenario

岡本 和憲[*1] OKAMOTO Kazunori
中野 有紀子[*2] NAKANO Yukiko
西田 豊明[*3] NISHIDA Toyoaki

[*1] 東京大学大学院情報理工学系研究科 Graduate School of Information Science and Technology, The University of Tokyo
[*2] 科学技術振興機構 Japan Science and Technology Agency
[*3] 京都大学大学院情報学研究科 Graduate School of Informatics, Kyoto University

Abstract: In virtue of great advances of computer graphics, the quality of CG contents has been getting higher and more real. Although CG contents are attractive and comprehensible for audiences, it is still difficult for non-professional users to create their own contents. In order to support the users to produce CG contents more easily, we have developed a system that automatically generates proper gestures for conversational agents and camera work based on a scenario. This system contributes to alleviating the difficulties of creating CG contents, and shows possible future directions for CG contents generation.

1. はじめに

近年の CG(Computer Graphics)技術の進歩により, CG キャラクタの外見的なリアリティは非常に高いものとなった. また, スクリプト言語を使用して, CG キャラクタを動作させることや, 映像ファイルや音楽ファイルを組み込む技術も確立し, CG コンテンツを制作することが身近なものとなった.

しかし, こうした CG コンテンツの制作において, 素人が思い通りのコンテンツを制作することは非常に困難で, ビデオコンテンツなどに比べ演出への工夫が乏しいことが多い. 演出の少ないコンテンツが多く制作される理由として, スクリプト言語による台本の作成は, 劇や映画といった抽象度の高い台本とは異なり, 演出の詳細な設定をしていく必要があり, 高度な専門知識やノウハウが要求されることがある. 特に, 近年の CG 技術の進歩による会話エージェントの外見的なリアリティの高さに見合う, ジェスチャなど, 動作の面でのリアリティを詳細に設定していくことは困難で手間のかかる作業である.

そうした背景を受け, 本稿ではニュース番組形式のコンテンツに注目し, 台詞を主体とした誰もが容易に作成できる台本から, CG コンテンツへの演出として, キャラクタによるジェスチャとカメラワークを自動的に付加するシステムを提案する.

2. 関連研究

本章では, これまでの主なＣＧコンテンツ生成の研究, ジェスチャ自動生成の研究について説明し, 本稿の目的を明確にする.

2.1 ＣＧコンテンツ自動生成の研究

ＣＧコンテンツ生成の研究例として番組記述言語 TVML(TV program Making Language)[林, 1996], MPML(Multimodal Presentation Markup Language)[筒井, 2000], Virtual Director[K. Manos, 2002]などがある.

本システムの最終的な出力のプラットフォームとして, 詳細なジェスチャなどを記述でき, 本稿が目的とするニュース番組形式の情報提供コンテンツの生成に優れた TVML を利用する. そこで, 本節では特に TVML について詳しく紹介する.

TVML(TV program Making Language)は, CG によるテレビ番組を記述するためのマークアップ言語であり, 実際のテレビ番組制作現場で用いられている番組台本の記述法に基づき, デザインされている. TVML では, 番組制作に必要な機能であるスタジオショット, スーパーインポーズ, タイトル, 動画ファイルの再生, オーディオファイルの再生, ナレーションなどが記述できる. こうして記述された TVML スクリプトを TVML プレーヤに送ると, CG コンテンツとして出力される. 具体的な TVML スクリプトの例を図 1に示す.

```
set: assign(name=info)
character: casting(name=MASA)
character: casting(name=MINA)
light: assign(name=no1)
camera: assign(name=Acam)
prop: assign(name=table)
…
…
prop: visible(name=picture1…)
camera: switch (name=Acam)
character: bow(name=MASA)
character: talk(name=MASA,text="こんにちは")
character: talk(name=MASA,text="今日のニュースを. . . ")
…
```

図 1:TVML スクリプトの例

図 1のＴＶＭＬスクリプトの例からわかるように, スクリプト言語による台本の作成では, キャラクタの動作や詳細な設定をすべて人手で記述する必要があり, それには高度な知識やノウハウが要求される.

2.2 ジェスチャ自動生成の研究

実世界で人が話をしている際のジェスチャを観察し, それに基づいてジェスチャを自動生成する研究例として E-COSMIC(Embodied Communication System for Mind Connection)[渡辺, 2003]や CAST(the Conversational Agent System for neTwork applications)[Nakano, 2004a]がある. E-COSMIC は実際の人間の音声情報に基づいて, ジェスチャを生成する研究であり, 会話の内容は考慮していない. そのため, 台本のようなテキスト情報からジェスチャを生成することができない. CAST は自然言語のテキストを入力とし, ジェスチャの提案

連絡先:岡本 和憲, 東京大学大学院 情報理工学系研究科, 〒113-8656 東京都文京区本郷 7-3-1, 03-5841-8758, kazu@kc.t.u-tokyo.ac.jp

を行う. しかし, CAST は仮想空間内のキャラクタや物の配置を考慮したジェスチャは生成しない.

2.3 関連研究のまとめ

以上, TVML では CG コンテンツの詳細なスクリプトを記述できるが, その作成コストは多大である. 一方, ジェスチャ自動生成の研究では, 空間の配置情報を考慮したジェスチャ決定への検討が不十分である. そこで本稿では, 簡単な配置情報と台詞からなる台本から, 配置情報を考慮したジェスチャ決定とカメラワークの決定を行い, TVML 形式のスクリプトを自動的に生成するシステムを提案する. 本研究では, CAST を拡張することにより, 配置情報を考慮したジェスチャ決定を実現する. 以下の章では, システムの詳細について述べる.

3. システムの概要

本システムの概要を図 2に示す. まず, コンテンツ制作者は台詞を主体とした台本を作成し, それを入力とする.入力された台本は3つの過程 (初期設定選択, ジェスチャ自動生成, カメラワーク自動生成) を経て, 最終的な出力として図 1で示した形式の TVML スクリプトに変換される.

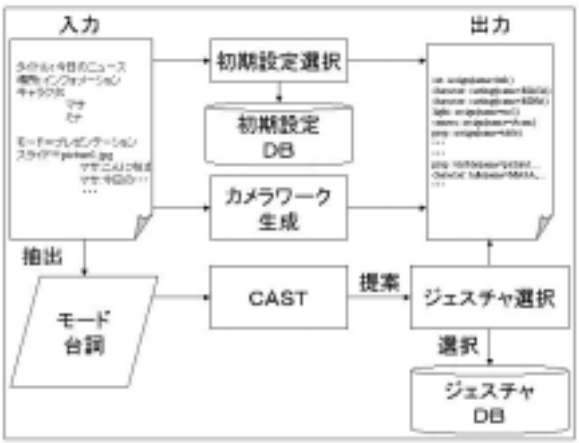

図 2:システムの概要

3.1 台本

システムへの入力となる台本には, 以下の情報が記述されている必要がある.

(1) 初期設定部分

TVML スクリプトで要求される初期設定には, セットやキャラクタのキャスティング, キャラクタやセットに必要な物(パネル, テーブル, 椅子)の配置, カメラやライトのセッティングなどがある. 本稿ではニュース番組形式の情報提供コンテンツに注目したので, コンテンツ制作者が記述する要素をタイトル, セット, キャラクタのキャスティングの3つに限定する.

(2) 本編部分

台本の本編部分は, 台詞, データファイル, モードの3つの要素からなる. データファイルとはパネルに映し出すスライドなどのことを指す. モードとは, 会話エージェントがニュースを読み上げるプレゼンテーションモードであるか, 会話エージェント同士が話をする会話モードであるかを記述する. 本システムにおいてモードは, 話している会話エージェントの視線に影響する. 具体的な台本の例を図 3に示す.

```
タイトル:今日のニュース
セット:スタジオ
キャラクタ:
        マサ
        ミナ

//本編
モード=プレゼンテーション
スライド:picture1.jpg
        マサ:このような流れで説明していきます。
        ・・・
```

図 3:台本例

3.2 初期設定選択

初期設定選択では, 台本に記述されたタイトルやセット, キャラクタを基に, 初期設定データベース(DB)を参照し, TVML スクリプトにおける初期設定を決定する. 具体的には, 台本に記述されたセットを設定し, 会話エージェントおよびスライドなどを映し出すパネル, 椅子やテーブルといった小道具をセット内に配置, また, それらの配置に適したカメラおよびライトのセッティング, オープニングでのタイトルの表示を設定する. 図 3の台本例からは図 4のような初期設定を生成する. この初期設定は 3.3(2)で説明するジェスチャ選択に影響する.

```
set: assign(name=info)
character: casting(name=MASA)
character: casting(name=MINA)
light: assign(name=no1)
camera: assign(name=Acam)
prop: assign(name=table)
...
super: on(type=infilehtml...
. . .
. . .
$infilehtml opening
<BODY LEFTMARGIN...
今日のニュース<BR>
<BR>
. . .
$end
```

図 4:初期設定の例

3.3 ジェスチャ自動生成

ジェスチャの生成は 2 段階で行われる. 第1段階は, CAST によるジェスチャの提案である. 次に, CAST で提案されたジェスチャと, 3.2節において選択したセット内の配置情報に基づいてジェスチャが選択される.

(1) ジェスチャ提案

まず, ジェスチャの提案を行う CAST について説明する. CAST はテキストを入力とし, エージェントの動作決定とそのタイムスケジュールの計算, およびエージェントの発話となる合成音声の生成を行う. CAST は(ⅰ)エージェント動作決定機構(Agent Behavior Selection Module (ABS)), (ⅱ)言語タグ付与機構(Language Tagging Module (LTM)), (ⅲ) エージェントアニメーションシステム, (iv)音声合成装置, の 4 つの主要構成素か

らなる. CAST に入力されたテキストは, まず LTM で言語情報のタグが付与される. この言語情報に対して, ABS がジェスチャ決定ルールを適用することにより, どの文節でどんなジェスチャを行うべきかが決定される. その結果, 表情やジェスチャ等のエージェントの動作タグがテキストに付与される. 最後に, テキストが音声合成装置に入力されることにより, 音声ファイルが作成され, それと同時に ABS は音素や文節区切りの時間情報を合成エンジンから取得し, これを基にジェスチャが実行されるべき時間を割り出し, エージェントアニメーションのタイムスケジュールを作成する.

次に, ABS の処理について詳細に述べる. ABS によるジェスチャ決定は, 日本語解析器[Kurohashi, 1994]による言語情報タグの付与と, タグつきテキストに対するジェスチャ決定ルール適用とからなる 2 段階のメカニズムで行われる. 以下それぞれについて, 詳しく述べる.

言語情報タグの付与:LTM では, 並列構造を含む文節間の係り受け関係, 新・旧情報, 助詞の種類 (e.g.,格助詞, 提題助詞), その他, 疑問詞, 強調の副詞, 数詞, 指示詞等についてタグ付けを行う. これらの項目は, 言語学研究と実際のデータ分析から抽出されたものであり, ジェスチャ決定への有効性が実証されている[Nakano, 2004b].

{テキスト ID:1, 文 ID:1, 文節 ID:9, 係り受け_from:8,係り受け_to:13, 文節タイプ:用言, 言語的分量: NA, 格: NA, WH 疑問: false, 新／旧情報: 新, 並列関係: 13, 強調副詞: false, Cue-Word: false, 数詞: false}

例えば上の例では, この文節の ID は 9 であり, 文節 8 がこの文節に係っており, この文節は文節 13 に係る. この文節は新情報を伝達し, 文節 13 と並列関係にある.

ジェスチャの決定:次に ABS は, 各文節に対し, ジェスチャを付与すべきかどうかを, ジェスチャ決定ルールを参照することにより決定する. 例えば, 先に示した文節は, 並列構造の構成素であるが, この場合には, システムは 47.7%の確率でジェスチャを該当文節に付与する (ジェスチャ決定ルールの詳細は[Nakano, 2004b]に譲る). 現在のシステムでは, ジェスチャの形態のデフォルトとしてビートジェスチャを採用している. ビートジェスチャとは, 手を上下に振るような身振りであり, 発話の意味内容とは直接的に関連せず, 発話の中で強調される部分に出現しやすい. 一方, 強調される文節中の概念に対して, 特定のジェスチャがエージェントアニメーションシステムのライブラリに定義されている場合 (例えば, 「大きい」という概念を表現するジェスチャがライブラリに既に登録されている場合) には, それが優先して用いられる. ジェスチャが決定されると, エージェント動作タグが XML 形式で付与される.

(2) ジェスチャ選択

ジェスチャ選択では, CAST で提案されたジェスチャ, 台本に記述されたモード, セット内の配置情報に基づいて, ジェスチャコマンドを選択する. 配置情報には方向の情報と距離の情報がある.

まず, モードに基づくジェスチャとして視線がある. プレゼンテーションモードでは, カメラに対して視線を送るか逸らすかのどちらかである. 会話モードでは, 視線はカメラか次の話者となる会話エージェントに送る.

次に, 方向の情報に基づくジェスチャとして, 指示動作がある. 指示動作は, 司会がパネルを指すような動作である. 初期設定で選択した配置情報を基に, 具体的な指示動作を選択する. 例えば, 司会の右側にパネルが配置された場合, パネルを指す指示動作は右手で右側を指す動作を選択する.

最後に, 距離の情報がジェスチャの選択にどう影響するかを説明する. 各ジェスチャには, そのジェスチャがどのくらいの空間を必要とするかを表すタグ情報があり, その情報と各キャラクタの大きさと掛け合わせた情報を基に会話エージェント同士, あるいはキャラクタとものが接触しないようなジェスチャを選択する.

ジェスチャ生成の具体的な処理としては, 図 3の台本例にある「このような・・・」という台詞から, 図 5のような詳細なジェスチャを含むスクリプトを生成する. g1, p5 は TVML のポーズコマンドで定義した具体的なポーズの形である. g1 は視線, p5 は指示動作の具体的なポーズの形である.

```
character: talkfile(name=MASA, filename=UID0, wait = no )
wait (time=0.1)
character: pose(name=MASA, pose=g1,wait=no)
wait (time=0.3)
character: pose(name=MASA, p5)
character: wait_talkfile(name=MASA)
```

図 5:詳細なジェスチャの記述

図 5の TVML スクリプトを TVML プレーヤに送り, CG コンテンツとして出力すると図 6のようになる.

図 6:出力の例

さらに本システムでは, ジェスチャを効果的に見せるためのアニメーションを生成することも可能である. 例えば, ビートジェスチャやコントラストジェスチャによって強調される語句の文字列を CAST から受け取り, これらの言葉のスーパーインポーズを入れることができる. これは, ジェスチャをする会話エージェントの位置, 選択されたジェスチャを基に, HTML タグを選択することにより出力される. 「A 社製品と B 社製品を比べると・・・」という台詞に対して, 本システムではコントラストジェスチャを生成するが, これにスーパーインポーズを入れた例を図 7に示す. 本システムは TVML スクリプトを CG コンテンツのスクリプト記述言語として採用しているが, CG の描画には, TVML プレーヤの使用は必須ではない. 本稿で提案した機構は, TVML プレーヤ以外のプラットフォームへも適用することが可能であり, プラットフォームを切り替えることにより, さらに多様なアニメーション効果を付加することができる[Li, Q, 2004].

3.4 カメラワーク自動生成

会話エージェントが複数の場合, カメラワークは会話エージェントのジェスチャを効果的に見せるために必要である. まず, カメラワークに必要なショットについて検討する. 実際のテレビ番組の1コーナー (「特ダネタイムズ」,1週間分) に多く見られたショ

ットを基に，本システムで用いるショットとして，以下の5つ提案する．それぞれのショットの例は図 8に示す．

(a) ロングショット（セット全体の配置を見せるショット）
(b) 司会のバストショット
(c) 司会とパネルの2-ショット
(d) コメンテータのバストショット
(e) パネルのアップショット

上述の5つのショットをつなぐカメラワークの生成に，図 9のショットの状態遷移モデルを用いる．

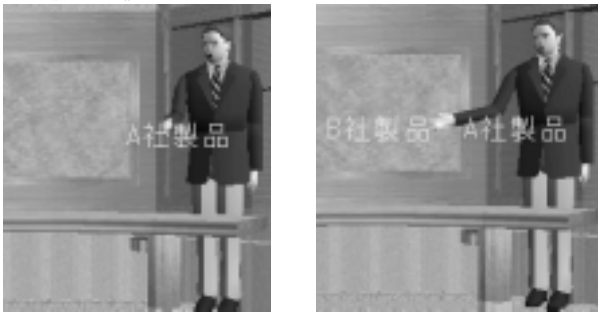

図 7:スーパーインポーズの例

図 8：ショットの例

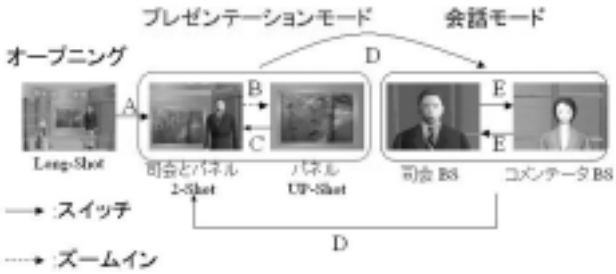

図 9：ショットの状態遷移モデル

遷移条件
A：時間（1～2秒）
B：司会の言語情報：「こちら」など
C：時間（5～6秒）
D：モードの切り替わり
E：ターンテイキング

各遷移条件の詳細を説明する．A はオープニングでロングショットを1～2秒見せて遷移させる条件である．B は司会者のパネルを指す（「こちら」などの）言語情報に基づいて遷移させる条件である．C はパネルを5～6秒見せて遷移させる条件で，D は台本に記述されたモードが切り替わった際の遷移条件である．プレゼンテーションモードから会話モードへの切り替わりの際は，司会あるいはコメンテータのうち話しているほうのバストショットにスイッチする．会話ーモードからプレゼンテーションモードへの切り替わりの際は，司会とパネルの2-ショットにスイッチする．E は会話モードにおいて発話者のバストショットを映すための条件である．

また，ショット間の遷移におけるカメラワークでは，スイッチとズームを用いる．

4. まとめ

本稿では，誰でも容易に記述できる台本から，会話エージェントの詳細なジェスチャやそれへのアニメーション効果，さらにはカメラワークを付加した CG コンテンツのスクリプトを自動生成するシステムを提案した．本システムにより情報提供を目的としたコンテンツ制作において，会話エージェントのジェスチャの詳細な設定をする労力を軽減することができ，誰でも簡単に CG コンテンツを作成することができる可能性を示した．

今後は，本システムの他のプラットフォームへの応用や，スーパーインポーズ等 CG コンテンツを効果的に見せるための工夫をさらに行っていく予定である．また，コンテンツ作成をさらに効率化するために，ジェスチャライブラリや配置情報の自動生成についても検討していきたい．例えば，現実の場面や人間の活動をコンテンツ作成に利用しやすいように記録することにより，CG コンテンツ作成のコストをさらに小さくすることができると考える．

参考文献

[林, 1996] 林正樹：テキスト台本からの自動番組制作～TVML の提案,1996 年テレビジョン学会年次大会,S4-3,pp589-592 ,1996.

[筒井, 2000] 筒井貴之, 石塚満：キャラクタエージェント制御機能を有するマルチモーダル・プレゼンテーション記述言語 MPML,情報処理学会論文誌, Vol.41,No.4,pp.1124-1133,2000.

[K. Manos, 2002] K. Manos, T. Panayiotopoulos, G. and Katsionis：Virtual Director: Visualization of Simple Scenarios, 2nd Hellenic Conference on Artificial Intelligence, SETEN, 2002.

[渡辺, 2003] 渡辺富夫：身体的コミュニケーションにおける引き込みと身体性ー心が通う身体的コミュニケーションシステム E-COSMIC の開発を通して, ベビーサイエンス, Vol.1.2, pp4-12, 2003.

[Nakano, 2004a] Nakano, Y., Murayama, T., and Nishida, T.：Multimodal Story-based Communication: Integrating a Movie and a Conversational Agent, IEICE Transactions, Special Issue on Human Communication (to appear), 2004.

[Kurohashi, 1994] Kurohashi, S., and Nagao, M.：A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. Computational Linguistics, 20(4), 507-534, 1994.

[Nakano, 2004b] Nakano, Y.I., et al.：Converting Text into Agent Animations: Assigning Gestures to Text. in HLT/NAACL 2004 (short paper). 2004.

[Li, Q., 2003] Li, Q., Nakano, Y., Okamoto, M., and Nishida, T.：Highlighting Multimodal Synchronization for Embodied Conversational Agent, Proceedings of the 2nd International Conference on Information Technology for Application, 2004.