物語のメンタルシミュレーションを生み出すメカニズムについて
From Text to Animation: Exploration of Mental Simulation during Story Text Comprehension

野口 武紘†，島崎 太一‡，榎津 秀次‡
Takehiro Noguchi, Yoshiki Komatsu, Hideji Enokizu

†芝浦工業大学大学院工学研究科，‡芝浦工業大学大学院工学研究科，‡芝浦工業大学
Graduate School of Engineering,Shibaura Institute of Technology, Shibaura Institute of Technology
m110107@shibaura-it.ac.jp, ma11082@shibaura-it.ac.jp, enokizu@shibaura-it.ac.jp

Abstract
In order to investigate the mechanism that underlies mental simulation during story comprehension, we designed the system that can generate animation from the story text. Mental simulation is analogous to mental reproduction of the microworld depicted by each sentence. It was revealed that animation needs information composing the microworld as well as information determining how to shoot the microworld. Based on these findings, we indicated the mental representation and some mental computations necessary to mental simulation.

Keywords — Event Indexing Model, Mental Simulation, Event Segmentation Theory, Automatic Generation of Animation, Film Language

1. 研究概要

人間が文章を読む際，文章に書かれている状況を頭の中でイメージし，そのイメージは，読者自身が文章に展開される状況に自分自身を置き，あたかもそこにいるようにイメージしていると言われている．

本研究はこの頭の中のイメージ（心的小世界）を構築する過程のモデルを参考にし，心的小世界に近いアニメーション映像を自動生成することで，人間の理解支援，あるいは文章を読んだときと同様に理解ができるようなシステムの構築を目指す．

これまでにも，文章からアニメーションを生成する研究はなされてきたが，それらの中でも，心的小世界の視点や，何に注目して構成されているか，といった人間の理解モデルの部分に焦点を当てた研究は十分にされているとはいえない．本研究では，心的小世界の視点というものは決して固定ではなく，注目すべき出来事やものが更新される毎に変わるものだと考える．そこで本研究は，人間が心的小世界をつくりだす際にどこに注目しているのかを考え，これらの視点の変化をアニメーションのカメラ操作としてシステムに組み込む．この操作を行うことで人間が頭の中でつくる状況にいるような没入感を与えるアニメーションを生成する．

2. 理解モデル

人間の物語理解は，テキストに記述された状況についてのイメージを構築する過程であり，そのイメージを心的小世界と呼ぶ．本研究ではテキストからアニメーションを生成するアルゴリズムを人間の理解モデルである Zwaan の Event Indexing Model と Zacks の Event Segmentation Theory の 2 つのモデルを導入する．

心的小世界を構築するために，まず人はテキストから表層的な言語情報を得る．次に，Zwaan の EIM(Event-Indexing Model)によると，心的小世界は主に時間，空間，人，物，因果，目標に注目し，それらの情報を統合し脳内で心的小世界をつくり理解しているといわれている．

本研究ではこのうち時間，空間，人物，物，そして目標をそこから起こる行為としてこれらに注目する．因果は自動生成に組み込むには難しく，また次に説明する物語の区切りによってある程度補えると考えたため，必要な情報として直接は導出していない．本研究におけるテキストから導出する情報のイメージを図 1 に示す．

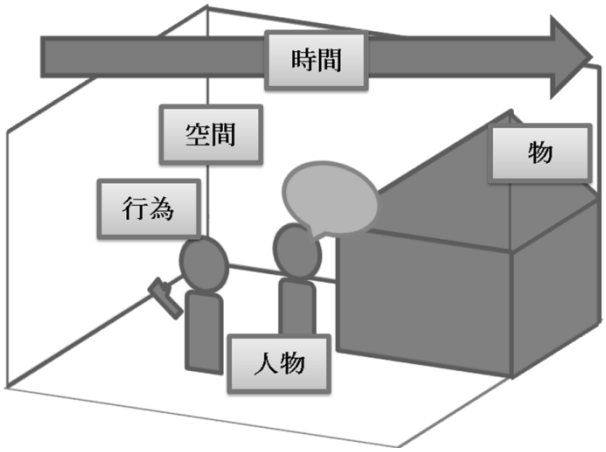

図 1　状況モデルの注目すべき要素のイメージ

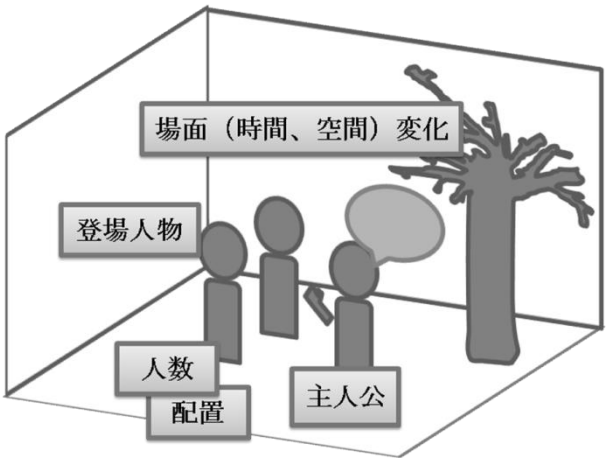

図 2　カメラ操作を決める要素のイメージ

Zacks の EST(Event Segmentation Theory)によれば，人間は物語内で起きるいくつかの特徴をとらえ分割し，意味の単位としてまとめ，複雑な動作のある世界を理解していると言われている．すなわち，人間は物語内の連続した出来事を特徴の変化毎に分割し，それらを意味のある単位毎に何が起きているか推論しているということである．特に EST によれば，時間，空間に注目し，その変化毎に物語を分割することで，我々はその内容を理解していると言われている．

本研究は，Zacks の EST を用いることにより，EIM によって導出された物語テキストに記述されている情報を意味の単位で分割し，構造化する．

人間が心的小世界を生成する際，自分自身がまるでテキストに記述された状況の中心にいるかのように感じ，物語に没入することで理解をしていると考えられている．つまり，常に遠くからその状況を傍観しているような視点の固定された舞台的な表現ではなく，映画のように注目する情報によって視点を様々に変化させていると考えられる．本研究では EIM や EST の理論による人間の物語理解を基に導出された情報を手掛かりに，視点の変化をカメラ操作として実装し，人間が文中のどの要素に注目しているかといった情報から撮影カメラを決定，より没入しやすいアニメーションの生成を行った．本研究におけるカメラ操作決定情報、すなわちテキストの情報から得られる計算，推論情報を以下の図 2 に示す．

3. システムの構成

本研究の大まかなシステム構成及び処理は以下図 3 のようになっている．

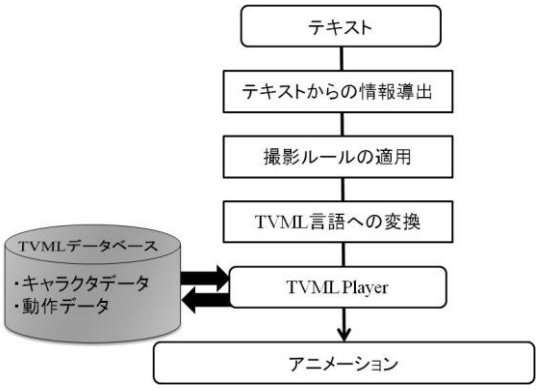

図 3　アニメーション自動生成システムの構成

本システムは，物語テキストを入力とし，まず自然言語処理を行い，一文ごとにアニメーション生成に必要な言語情報を導出する．解析から得られる単語の情報をフレームに格納し，更新と計算を行う．次に本研究におけるアニメーションを出力するうえで重要なカメラ操作を決定するため，これらの情報に撮影ルールを適用する．以上の過程で導出したアニメーションに必要な情報を全て TVML 言語に変換し，TVML Player に読み込ませる．また，この際に TVML に内蔵されているデータでは足りないキャラクタや動作のデータをデータベースから引用し，最終的にシーンごとに分割されたアニメーションが生成される．

4. テキストからの情報の導出

初めにテキストからアニメーションに必要な情

報を導出する処理について説明する．この部分の構造は図 4 の通りである．

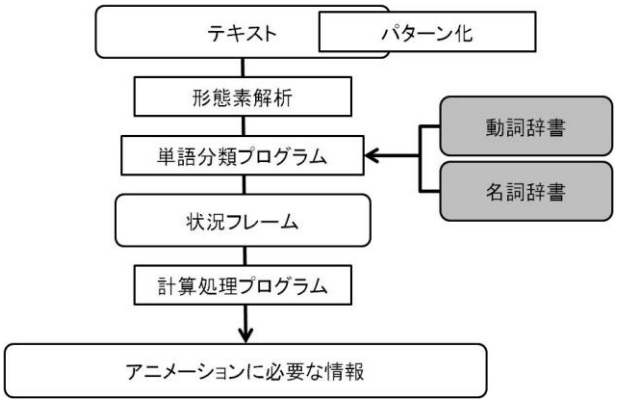

図 4　アニメーションに必要な情報の導出過程

本研究では，入力となるテキストを解析しやすいようにあらかじめパターン化した状態で扱う．このパターン化によって，一文に行為は一種類としている．まずテキストに形態素解析を施し，単語に分割する．本来ならばここで行うべき構文解析は文のパターン化によって必要がなくなり，意味解析は現段階では実用的な処理が困難なため，テキストに出てくる単語の意味をあらかじめ辞書として用意することによって回避している．これらはよく用いられる名詞辞書と動詞辞書に分かれており，例えば動詞辞書には動詞ごとの入退場に関する情報といった，本研究に必要な情報も付加できるようになっている．

次に単語分類プログラムによって単語ごとの文中での意味を格助詞と主語から決定する．ここでは文中の名詞を深層格[3]に当てはめ，後の処理で利用しやすくしてある．また主語は，動作，感情，状態の全てを持つ「人物」，感情は持たない「物」，どちらにも当てはまらない「その他」に分類した．この決定のアルゴリズムは表 1 のようになる．

表 1　格助詞と主語による深層格の決定

格助詞＼動作主	人物	物	その他
が/は	動作主		
に	被動作主	対象	空間
へ	被動作主	対象	終点
を	被動作主	対象	空間
、	—	—	時間
から	—	—	起点
で	—	道具	空間

ただしいくつかの例外があり，例えば従属や装飾を示す格助詞「の」については状況によって前後の単語の関係が大きく変わるため，現段階では処理方法を決定しきれていない．また表には記載していないが，助動詞「と」については常に並列を表す助動詞として扱い，これは前後の単語の深層格を等しくする働きをもつ．

このようにして決定された文中の各単語は，その役割に応じて状況フレームに格納, 編成される．

4. 1. 状況フレーム

本研究では，Zacks の EST と Zwaan の EIM を用いた人間の談話理解モデルの知識表現に Minsky のフレーム理論を用いた．これを状況フレームと呼ぶ．

一般的にフレームは階層構造の形を取る．EIM 及び EST もまた, 必要とする情報によって重みがあることを考えると階層構造と言えるため，これらは親和性が高い．各フレームは，そのフレームが持つ要素を表すスロットを持ち，スロットには状況に応じて様々な値が代入される．スロットに値が代入されると，そのスロットに関連する情報を表す下位フレームを探索、移動し，下位フレームが見つからなかった場合は新たに生成される．ただしこれを繰り返すと無限にフレームを生成しなければならなくなるため、本研究では状況モデルにおける重要な情報をそのままフレームとし，それ以上に細かい情報のフレームは生成しないものとする．本研究における状況フレームの構造を図 5 に示す．

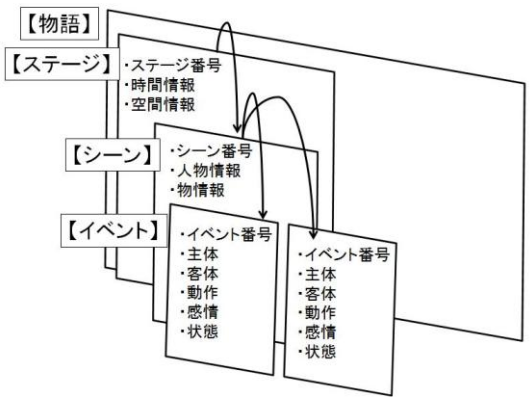

図 5　状況フレームの構造

解析対象の文における，最も重みが大きいと考えられる時間，空間の情報をステージと呼び，このステージフレームが最初に生成される．ステージフレームは最も重みのある情報群であり，ステージ内にある人物と物に関わる情報を内包している．次に空間の中に存在する人物，物に関するフレームが生成される．同じ時間の流れかつ，同じ空間での人物情報であることからそれをシーンフレームと呼ぶ．最後にシーン内で表現される人物の動作についてのフレームが生成される．これをイベントフレームと呼び, 主に人物の発話, 動作, 感情についての情報が格納される．また，人物の発話あるいは会話は他の行為に比べて情報の重みがあると考え，本研究では会話に関するフレームを実装した．ただし会話内容を解析すると非常に複雑な処理が必要になってしまうので，フレームの生成はここで終了となる．

4. 2. 推論情報の導出

テキストから直接状況フレームに編成した情報を用いることで舞台設定は整うが，本研究ではメンタルシミュレーションにおいて注目する場所がどこであるかが，メンタルシミュレーションを構成する重要な要素と捉える．そこで状況フレームに編成された要素に計算，推論を加え，アニメーションにおける視点の変更すなわちカメラ操作を決定するために必要な推論情報を導出する．これらは複数の要素や複数の文から決定する．表 2 にその種類と性質を示す.

表 2　カメラ操作を決定する推論情報

ステージ変化の判断	ステージが変化した文の推論と状況説明ショットの選択
既存人物とその人数	シーン変化の前後で変わらず登場する人物とその人数の判断
新規人物とその人数	シーン変化によって新たに登場した人物とその人数の判断
人物の登場回数	文章全体及びシーン単位での登場回数を計算、重要度の決定
主体、客体の判断	動作の主体及び客体を文から推論し、撮影する人物を決定

まずステージ情報の変化が起こることで，状況説明ショット(establish shot)の利用を判断できる．また，本研究におけるカメラ操作は，基本的にシーンにおける人物の配置，動作，及びカメラの配置といった情報から決定する．人物配置は，シーンごとの新規，既存人物の人数や登場回数から決定でき，カメラで何を映すかについては，現在起こっている動作の種類，動作の主体や客体といった情報から決定できる．このような追加情報によってアニメーションがより詳細に決まる．

状況フレームと推論情報が全て編成、決定されると，これを状況テーブルと呼ぶ CSV 形式のテキストファイルにして一度出力し、撮影ルールの適用を行い最終的にアニメーションに必要な情報を導出する。

5. 撮影ルールの適用

これまでに導出されたアニメーションに必要な情報から if-then ルールによりカメラ操作を決定し，生成されたアニメーションに状況によって変化する視点の概念を導入する．撮影ルールの適用及び最終的なアニメーションの出力までの過程を図 6 に示す.

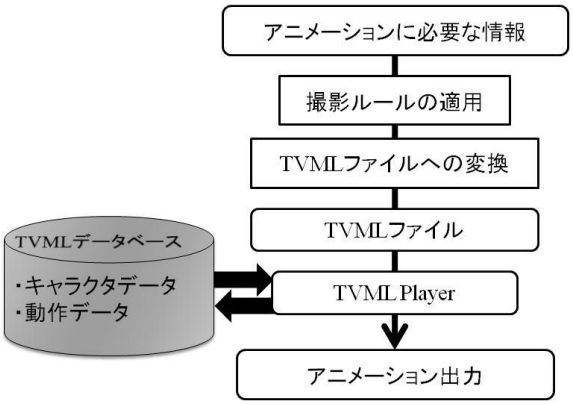

図 6　最終的なアニメーション生成までの過程

今回，視点を決定する撮影ルール構築のために映画文法[6]を利用した．映画文法とは映画制作関係者によって経験則的に養われた知識を自然言語により記録したものである.

状況テーブルを読み込み，文ごとの情報に撮影ルールを適用する．決定したカメラ操作の情報を加え，これらの情報を TVML[7]言語で記述された TVML ファイルに変換する．これは本研究ではアニメーションの出力に TVML Player を用いるためである．TVML は独自のインタプリタ型言語である TVML 言語を持ち，TVML 言語で記述され

たファイルを TVML Player に読み込ませることでアニメーションを生成することができるソフトウェアで, TVML Player による映像はカメラを非常に自由に配置することができる．また本研究では, あらかじめ TVML に用意されたデータでは汎用性に問題があると見て，新たにキャラクタや動作のデータを作成し, TVML Player で利用できるようにしている．

5.1. 撮影ルールのカメラ操作モデル

本研究のカメラ操作モデルは，初めに場面情報が変わったかという情報，次に文中の動作主が人物または物であるという条件の下で，人物の配置ルールとカメラ決定ルールによってショットを決定する．このカメラ操作モデルを図 7 に示す．

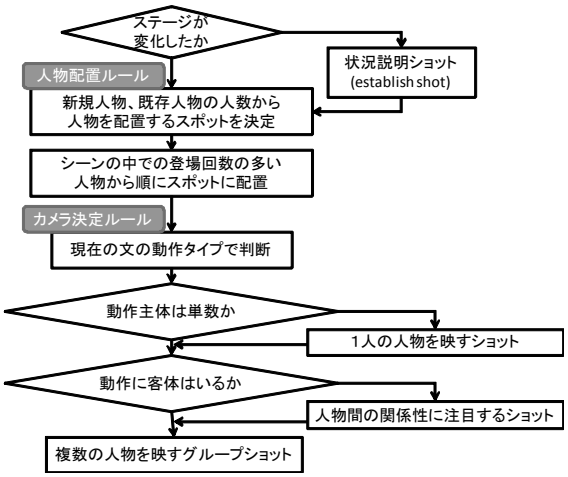

図 7　ショット決定のカメラ操作モデル

以下で各ショットの説明をする．

i.状況説明ショット(establish shot)

状況説明ショットとは Griffith[8]が提唱したものであり，場面の冒頭に挟むことにより視聴者が状況を理解しやすくなるショットのことをいう．後に説明する Long Shot がこれに該当する．

ii.　一人の人物を映すショット

このショットでは，注目する人物が画面の中央に収まるようにカメラを配置する．また，その人物の動作の種類を分析することでカメラのショットサイズを決定する．ショットサイズとは映像を意識的に空間文節するものであり，一般的には人物の心理的な距離の描写等に利用される．

本研究で扱う CG キャラクタは等身が現実世界の人間とは異なるため，図 8 のように統合した四種のショットサイズを利用する．

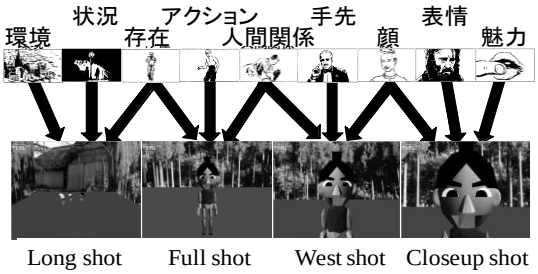

図 8　意味的な分類によるショットサイズ

iii.人物の関係性に注目するショット

会話や物の受け渡しといった，二人の人物の関連性を表すショットでは，個々の人物ではなく動作の受け手の表情や周囲の背景など，そのシーンの出来事そのものを映す必要がある．これを満たす撮影方法として求心ショット，中でも図 9 のように，一般的に映画製作等の会話シーンで扱われる肩越しショット(over-the-shoulder)，すなわち一方の人物の背後から頭と肩のリア映しこみ，その向こうに相手を映すといったショット等で撮影する．

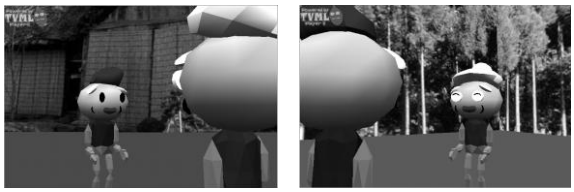

図 9　二人の人物の会話シーンの撮影

iv.より多くの人物に注目するグループショット

映画等でグループ撮影を行う場合，グループ内には必ず最も注目されるリーダー的人物すなわち主人公が存在し，その人物を中心にしてシーンが構成されるため，その人物を主体としたカメラ操作が行われる．そこで，物語あるいはシーンにおいて最も言及されることが多かった人物を中心に撮影するショットを扱う．

例として，我々が“桃太郎は犬に吉備団子をあげた”という文を読んだとき，撮影ルールに従うと動作主である桃太郎と被動作主の犬という二人の登場人物がいることから“人物間のインタラクションに注目するショット”によってこの状況をシミュレートする．このとき登場人数から二人は空間の中心で向きあって立ち，初めに被動作主で

ある犬の後ろから肩越しショットで撮影し，次に物を受け取った犬を桃太郎の背後から肩越しショットで撮影する，といったカメラ操作が決定される．

6. 結果

本研究の提案するシステムにより，アニメーションを生成してみたところ，我々の目指す違和感のないアニメーションが生成された．動作毎のカメラの切り替えがアニメーション内での出来事を捉えていることが大きな要因であると考えられる．しかし，１文について１つの動作，アニメーションに配置されているキャラクターの位置，これらは実際の映画から導き出したものではない．本来の映画であれば，例えば「食べる」といった動作はいくつかの動作からなり，それに対応した写し方がある．映画のようなアニメーションの生成を目指すのであれば，テキストからのより多くの情報を導出するアルゴリズムと，映画を解析し映し方や人物の配置を決めることが今後の課題である．

7. 考察

出力されるアニメーションは状況や人物の変化によって物語中で何が起きているのかを物語テキスト中から抽出し，アニメーションにおけるキャラクターや映し方を決めている．しかし，人物は新規の人物か，既存の人物か，出てくる人物が主体か客体化か, 登場回数によって決まる人物の配置，といった処理は人間の理解モデルに基づいた処理とは言えない．アニメーションを出力するための手掛りを導きだす処理であるため，今後，アニメーションの出力方法を変更する場合は，改善してゆく必要がある．

さらに，本研究で取り扱ったキャラクターの配置は映画文法を基にしているが，映画文法そのものは 1980 年代に作られた映画を統計的にまとめたものであるため，現代の映画と比べると必ずしも映画文法が正しいとはいえない．今後は，現代の映画を解析することにより分かりやすい映像を考える必要がある．

8. 結論

本研究は, テキストを入力としたアニメーション自動生成システムの構築に取り組んできた．今回の試行とシステムの構築により，アニメーション生成に必要な情報と，撮影に必要な情報，これらの２つの方針を決めることができた．

同時に，物語理解モデルを取り入れたアニメーション生成システムは，テキストからアニメーションを出力することで，物語理解過程のシュミュレートし，どのような情報処理過程であるのかを明らかにしていると言える．

参考文献

[1] Zwaan R.A and Radvansky G.A, (1998) “Situation Models in Language Comprehension and Memory”, Psychological Bulletin, vol. 123, No. 2, pp162-185.

[2] Zwaan R.A, (1999) “Situation Models: The Mental Leap Into Imagined Worlds”, Current Directions in Psychological Science, Vo.8, pp15-18.

[2] Zacks,J.M., & Kurby,C.A. (2008) “Segmentation in the perception and memory of events:”, Trends in Cognitive Sciences, 12(2), pp72-79.

[4] 長尾真，佐藤理史，黒橋禎夫，角田達彦，（1996） “意味解析”，自然言語処理，pp122-132, pp200-204.

[5] マーヴィン・ミンスキー，安西祐一郎（訳）（1990）“言語フレーム”，心の社会，pp421-443, pp553.

[6] Minsky M, 白井良明，杉原厚吉（訳） (1979) “A framework for representing knowledge”, コンピュータビジョンの心理, pp237-330.

[7] D. Arijon, 岩本憲児，出口文人（訳），(1980) 映画の文法.

[8] 林正樹，(2005) めざせ！テレビ番組クリエイター―パソコンと番組記述言語 TVML で実現!!―

[9] 純丘曜彰，エンターテイメント映画の文法