2004－HI－111　(19)
2004／11／12

実画像ベースによる手話映像合成の試み

鈴木　雄介†　竹内　晃一†　宮本　一郎†　三樹　弘之†

概要

手話を行う人物を撮影した実映像素片を連結、合成して手話文章映像を表示する手話映像合成表示システムを試作したので報告する。今回の試作では、気象概況を題材とし、使用頻度が高い約２００の手話単語の映像素片を組み合わせて、テキストで入力される気象概況を手話として合成・表示できるようにした。この試作システムでは、実映像素片を連結する際に、先行する映像ファイルと、後続する映像ファイルの中間部分で発生する人物の動きのギャップによる違和感を最小限にするため、モーフィング等の映像処理技術を用い中間部分の映像を合成し、動きを補完する処理を行なうことを試みた。

Real Image-based sign language synthesizing system

Yusuke Suzuki†　Koichi Takeuchi†　Ichiro Miyamoto†　Hiroyuki Miki†

Abstract

This paper describes a sign language synthesizing system. The system utilizes video captured deaf person's signing images for synthesizing sentences in sign language. We took sentences from daily general weather overview. By combining signed words sampled from these sentences, the system synthesizes and represents daily general weather overview in sign language.

We applied morphing techniques for synthesizing intermediate images to reduce unconformities found in transition parts of two images.

1. はじめに

近年、手話工学と呼ばれる研究が盛んであり、手話使用者のコミュニケーションの円滑化を目指した研究が進められている。その中でも、日本語を手話へ翻訳する手話翻訳システムの実現を目指し、手話映像を表現するためのシステムの研究は数多い。

これらのシステムの研究では手話映像の表示方法としてコンピュータグラフィックス(以下 CG)を用いているものが多い。これは手話単語間の遷移の表現や、動作データの加工が容易であるなどの利点があるためである。代表的なものとして、黒川ら[1]、黒田ら[2]による研究試作や日立製作所によって商品化されたソフトウェア[3][4]がある。

しかし、現在の CG による手話表示システムでは、手話において重要な役割を果たす表情要素、口形等の非手指要素の表現が不十分であると言われている。筆者らによる先行研究[5]では手話使用者、特に手話を母語とするろう者はＣＧよりも実写による手話映像表現を好むとの結果を得ているが、非手指要素の有無もこの差異の一因であろう。

手話における非手指要素とりわけ表情要素の重要性は手話分析の研究では以前より認識されている[6][7]。表情要素を文法要素として理解するろう者にとっても、口形から音を読み取る必要がある中途失聴者、難聴者にとっても表情要素は重要である。

表情要素に着目した CG 利用の手話表現の研究として 児玉、安村による研究[8]などがあるが、未だ課題は多い。

一般的な CG 利用の利点は映像加工、合成の容易さにあるが、表情要素までも合成しようとすると CG では困難である。これは CG で表情を表現するためのパラメータの次元が多いということだけでなく、手話における表情要素の分析が未だ途上にあり、表情要素の役割、表出方法を理解し、合成応用可能な形にすることが困難なためである。このような現状を鑑みると、CG によって、表情要素をはじめとした非手指要素を忠実に再現することは困難であると言える。

本研究では、このような表情要素や口形等の非手指要素の表現に着目し、CG を利用せず、ろう者による日本手話発話を撮影した実映像を用いた手話映

† 沖電気工業株式会社　研究開発本部　ヒューマンインターフェースラボラトリ
Oki Electric Industry Co., Ltd. Corporate Research and Development Center, Human Interface Laboratory

像合成を試みる。実写映像であるので、表情要素をはじめとした非手指表現もそのまま表現できることが特徴となる。

実写映像の表現力は高いが、任意の文章を合成することは CG に比べれば困難である。本論ではこの問題点を解決し、表情要素や口形の表現力の高い手話映像表現システムを作成することを目標とする。

2. 試作システム

2.1 システム構成

試作したシステムの構成の概念図を図 1 に示す。このシステムは入力された文章を形態素解析して、分割された各日本語の単語に相当する手話映像ファイルをデータベースから選択し、連結して一連の文章として表示するための構成である。

以下、図 1 に基づき、各要素とその実際の動作について説明する。

ユーザは文章入力部を通して表示したい日本語の文章を入力する。入力された文章は日本語形態素解析システム Chasen[9]を用いて形態素解析される。

解析結果は翻訳部に送られ、翻訳部は分割された単語から助詞等手話に単語として直接には表現されない部分を削除し、係り受けを調べ適切な単語名を

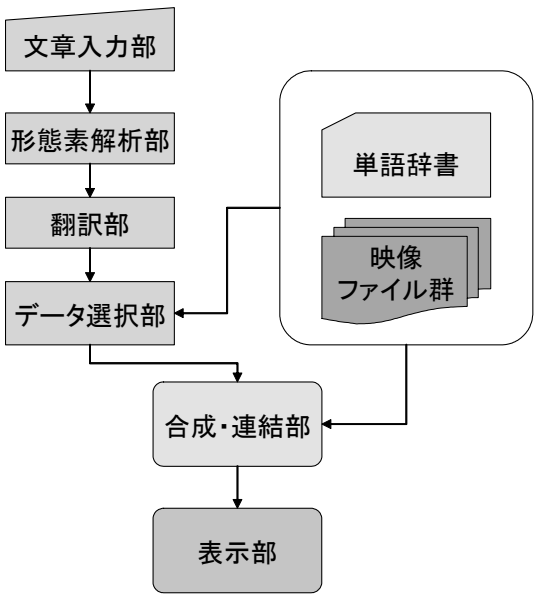

図1 システムの概念図

選択するなどの翻訳処理をする。

処理された品詞などのデータがついた単語を基にラベルを作成して単語辞書と呼ぶデータベースを検索、単語ごとに単語を表現するための映像ファイルを選択して映像ファイル群から抽出する。

映像ファイル群とは実際の映像を表示するための素材となる手話を表現している人物を撮影した映像ファイルをおおよそ単語単位で分節化し、ラベルをつけたものである。映像ファイル群の作成方法については次節で説明する。

選択された映像ファイルを用いて、合成部は映像ファイルの中に現れない人物の動作を合成する。

選択された映像ファイルと合成映像ファイルを連結し、最後に連結された映像ファイルを表示部で表示する。

2.2 表現する対象分野

今回の実験では、手話で表現する題材として気象概況を選択した。これは以下の条件を検討した結果である。

(1)気象概況は比較的定型的な文章が多く、限定された語彙で多数の表現が可能である。また複雑な文が少ないため日本語文章解析処理が単純化できる。

(2)気象概況には台風警報など緊急性の高い情報を伝えるという役割がある。緊急時には誰もが混乱し、情報の理解力が低下する。手話使用者の中でも先天性のろう者、幼児失聴のろう者は日本語と日本手話の多言語使用者である場合が多いが、緊急時にはろう者にとっての母語である手話での情報伝達が最も理解が容易なものになると考えられるため、気象概況を手話で表現することが、ろう者にとって重要なこととなる可能性が高い。

これらの理由から本実験では気象概況を日本語から手話の表現に変換するシステムの作成を目標とすることとした。

3. 撮影内容の説明

本節では合成に使用する素材である映像素片ファイルを撮影した環境、方法と、撮影された映像素片ファイルの内容について説明する。

3.1 撮影環境

使用した手話映像はネイティブサイナーによる手話映像をデジタルビデオカメラで撮影したものである。 シャッタースピードは 60[fps]で、手話動作を撮影してもビデオの各映像フレームに手のブレなどが生じず、後の映像合成処理が容易になるようにした。

人物の背景はクロマキー処理用ブルーバックとした。これは撮影する人物の位置あわせなどの事前の処理、また後の合成処理を容易にするためである。

3.2 撮影内容

撮影する内容は気象庁発表の気象概況の例文を解析して決定した。なお、1 回分例文は 5～6 文章前後からなり、200～300 字前後である。約 30 回分の例文を集め、形態素解析及び単語単位分割し、単語出現頻度順に 200 単語を選択した。

実際に手話と日本語の単語が対応するものを 165 個選択し、日本語の発話（口形）と同時に単語ごとに手話を聴覚障害者が表現したものを撮影した(単語レベルの手話映像) 。本研究では表情要素や口形の表示を重要視しているため手話の手指要素が同じで口形のみが異なる単語を別の単語として扱う。

さらに、選択した単語を用いて気象概況の文章を実際の気象概況の例文の単語を置き換える、数字を変えるなどの変更をした例文を約 50 文撮影した(文レベルの手話映像)。これらの文レベルで撮影した手話映像は、後の比較、解析などの参考として用いている。

3.3 映像ファイルの分節方法

手話動作の意味をもつ単語の部分と意味を持たない動作の部分をどのように区切るかは議論の続いている課題であるが、本試作システムにおいては、以下で説明する簡便なルールに従って、撮影した映像ファイルを単語単位に区切った。

まず図 2 に示す、手を重ねた状態を撮影時の基本開始姿勢とし、すべての撮影はこの姿勢の状態から開始するものとした。

図2 手話動作の基本開始姿勢

この状態から重なっている両手がはじめて画面上で二つに分かれたフレーム、または画面上での両肘を結んだ直線上より両手が上部移動したフレームの早い方を単語の開始位置とした(図 3 参照)。

二つに分離した　　肘間直線の上に移動した

図3 単語の区切り方法

終了位置は、手話発話終了時に約 9 割の率で入るまばたきを指標とした。完全に目を閉じる一つ前のフレームを終了位置とした。まばたきが確認できない場合は発話（口形）の終了時点を終了位置とした。

3.4 係り受け

日本手話と日本語は異なる言語であるため日本語では一つの単語で表現される内容が複数種類の手話表現となることがある。

これを気象概況に用いる単語を例に説明すると、図 4 に示すように「台風が北上する」の「北上する」と「前線が北上する」の「北上する」は形態が異なる。

「北上する」(台風)　　「北上する」(前線)

図4 係り受けによる形態の違い

また「数字」は係る単語が「台風」であるか「気温」であるかで左手(弱手)の位置が大きく異なる(図 5 参照)。

5号(台風)　　5度(気温)

図5 係り受けによる形態の違い

係りをうけて変化するそれぞれの場合ごとに一つのファイルとして、翻訳の際には形態を変える要因となる単語が文中に存在する場合に適宜ふさわしい映像ファイルが選択されるようにする。

4. 問題点

映像ファイルはほぼ単語ごとに分節化されたものであるので、これを単純に連結して表示すると各単語の終了時から次の単語の開始時までの人物の動作が最終的な連結映像に現れず、人物動作の空白域（ギャップ）が生じる。これによって最終的に表示される映像ファイルにおいて繋ぎ目が目立ち、違和感や見にくさを感じるという問題が発生する。

5.映像合成

ファイル間の繋ぎ目に発生するギャップを解消するため、システムの合成部はそれぞれの映像ファイルを解析しそのデータを用いてギャップを埋めるための映像を合成する。合成された映像を映像ファイルの前後の間に挿入することで最終的に表示する映像ファイルを作成する。結果最終的に表示される映像には単語間の遷移部分が補完され、繋ぎ目が目立たなくなり、違和感無く手話情報を伝達することが可能になる。

実際の映像合成処理について図 6 を用いて説明する。

1.データ選択部によって選択された映像ファイルから二つのファイルを抜き出す。先に表示されるファイルを先行ファイル、次のファイルを後続ファイルとする。

2.先行ファイルから動作の終了時点のフレーム（最終フレーム）を取り出し、後続ファイルから動作の開始時点のフレーム（先頭フレーム） を取り出す。

3.取り出した二枚のフレームに対して映像合成の技法であるモーフィング処理を行うことによって、最終フレームから先頭フレームへと変化する中間の映像フレームを複数枚生成する。

4.生成した映像フレームを連結して映像ファイル作成する。

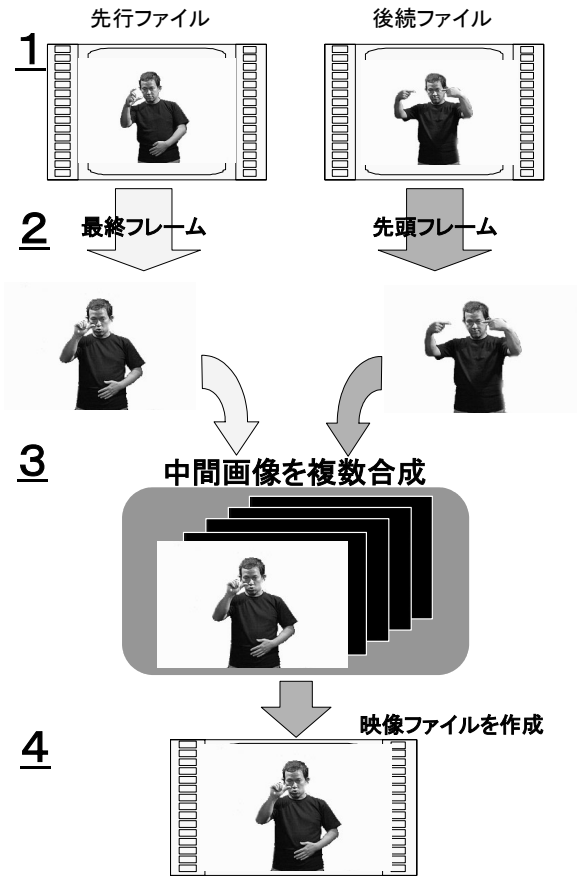

図6 映像ファイルの合成方法

こうして作成された先行ファイルと後続ファイルの中間部分に挿入する。モーフィングの手法として、クロスディゾルブ、メッシュワーピング、特徴ワーピングの三種類を用いた[10]。

前記二種類のワーピングでは合成中に変化する部分の画面内での領域や特徴線を指定する必要がある。手話映像で動作するのは主に腕領域であるため肌色部分抽出などの方法を用いて腕領域を映像処理により自動的に抽出することも可能である。しかし、今回は試験的な構成であるため、映像フレームの切り出し時に手動で領域、特徴線を指定しておき、そのデータを合成の際に参照して用いることとした。

三種の合成方法とも二枚の映像から、5 枚の中間映像ファイルを作成し、15[fps]の映像を合成する。結果 300[ms]の時間の中間映像が挿入されることになる。これらのパラメータは試行錯誤と聴覚障害者に対するインタビューによって決定した。

6. 結論

本研究では手話使用者にとっての利便性を高めるため、表現力の高い実写映像を利用し、単語間の動作を映像合成の手法を用いて補完する手話映像表示システムを提案、試作した。

システムの有効性については定量的な評価は未完であるが、手話使用者に対する予備的な聞き取りによれば、仮説通り、非手指要素特に表情要素や口形が表現されていることが読み取りやすさ・見易さに貢献しているらしいことが確認できた。合成に伴う結合部のギャップについても、手話読み取りに支障がある程ではなく、改善の余地はあるものの実用上は問題がなさそうであることが確認できた。

7.議論

本論文での試行はまだ初期段階のものであり、検討すべき未解決事項は数多いが、ここでは次の２点に絞って議論する。

・手話映像の分節について

今回の試作においては 3.3 節で述べたような分節を行ったが、実際の手話においては、前後の単語との関係で、明確な分節が難しい場合がある。具体的には前後の単語によって動作の開始・終了位置が変化する、前単語の動作がしばらく残る、次単語への予備動作が入るといったことが起こる。これらの音声言語で言えば「リエゾン」に相当する部分をどう扱うかについては、検討の余地がある。

・映像合成について

様々な映像合成方法を試みているが、手話のような意味のある身体動作の映像断片を繋ぐ場合には、人間の知覚特性や錯覚を利用するのがよさそうであることがわかってきた。必ずしも映像をすべて作成しなくても、人間側の予期や知識で補間映像があると認識される場合がある。このような現象をうまく利用すれば、複雑な映像合成処理をせずとも認知上自然な合成が実現できる可能性がある。このような点では、アニメーション製作で使われる各種テクニックが有効であると考える[11]。

8.今後の予定

本論執筆時点ではシステムの評価が十分に行われていないが、まずは合成される映像の評価を通してこのシステムを評価する必要がある。
この際には、システムの使用者として想定される手話利用者を対象とした評価実験が必須となる。これについては、質問紙による感性評価や手話内容の読み取り実験を実施予定である。一般に映像圧縮方法の評価等でも同様の評価法が採られるが、手話は表現する内容をもつ言語であるため、それにふさわしい評価方法を検討する必要がある。

映像合成方法の改良も進める。今回は合成される映像ファイルの表示時間を一定としたが、単語同士の結びつきに応じて時間を変化させるなどの工夫、また7章で議論したような人間の知覚特性を利用した方式改良が必要と考えられる。

今後も評価実験や方式改良を継続し、手話使用者にとって見易く、わかりやすく、利便性の高い手話映像合成表示システムの実現を目指したい。

謝辞

本研究を行うにあたり、御意見御指導並びに撮影機材使用にご協力頂いた工学院大学長嶋裕二教授と撮影にご協力頂いた住田英之氏に感謝いたします。

参考文献

[1]黒川隆夫,"手話と日本語の相互翻訳の試み" 第5回関西情報関連学会連合大会論文集,pp.25-34,(2000)

[2]黒田知宏, 佐藤宏介, 千原國宏,"手話伝送システム S-TEL", 電子情報通信学会技術研究報告,ET96-85,pp.65-71,(1996)

[3]佐川浩彦,"手話アニメーションソフト Mimehand とその応用", 医療とコンピュータ Vol.13,No.8,(2002)

[4]崎山朝子, 大平栄二, 佐川浩彦, 大木優, 池田尚司,"リアルタイム手話アニメーションの合成方法," 電気情報通信学会論文誌D－Ⅱ,Vol.J79-D-Ⅱ,No.2,pp182-190,(1996)

[5]市川貴士, 宮本一郎, 鈴木雄介, 竹内晃一,"携帯端末画面での手話映像の見易さに関する検討",ヒューマンインターフェースシンポジウム 2003 論文集,pp309-312,(2003)

[6]木村晴美, 市田泰博,"はじめての手話",日本文芸社,(1995)

[7]米原裕貴, 長嶋祐二,"手話の習熟度による注視点の変化に関する実験的検討", ヒューマンインターフェースシンポジウム 2002 論文集, pp233-236,(2002)

[8]児玉哲彦, 安村通晃, "表情の表現を含む手話アニメーションの試作", 情報処理学会研究報告,2003-HI-103,pp23-29,(2003)

[9] 日本語形態素解析システム Chasen, http://chasen.naist.jp/hiki/ChaSen/

[10]Thaddeus Beier , Shawn Neely, "Feature-based image metamorphosis", Proceedings of the 19th annual conference on Computer graphics and interactive techniques,pp35-42,(1992)

[11]Ollie Johnston, Frank Thomas, The illusion of Life: Disney animation, Disney Editions, 1995