

姿勢推定を用いた手話映像からのキーフレーム候補抽出

丸亀 泰作^{†1} 米村 俊一^{†2} 筒口 拳^{†1}

概要: 手話映像の効率的な要約のため、手話映像中から手話の特徴を強く表している画像（キーフレーム）を抽出してキーフレームだけから映像を再構成する方式を検討している。その中でキーフレームをどのように自動抽出するかが大きな課題である。本研究では、キーフレームでは手指の運動の軌跡が停留値を取るという仮定のもと、映像中の手話者の姿勢推定により手の関節位置の時間変化を曲線近似し、その曲線が極値をとるフレームをキーフレーム候補として抽出を試みた。実験の結果、正解キーフレームの8割程度を自動抽出できることがわかった。

キーワード: 手話, 映像要約, キーフレーム, 姿勢推定, 曲線近似

Extraction of Keyframe Candidates from Sign Language Video using Posture Estimation

Taisaku MARUGAME^{†1} Shunichi YONEMURA^{†2} Ken TSUTSUGUCHI^{†1}

Abstract: For efficient abstraction of sign language videos, we are studying a method of extracting images called keyframes that strongly represent the characteristics of sign language from the videos, and reconstructing the videos from only the keyframes. The major issue is how to automatically extract keyframes. In this study, based on the assumption that the locus of hand movement take a stationary value in keyframes, the time change of the joint positions of the arm of the person are approximated by a curve approximation by estimating the posture of the hand. We tried to extract the frame that the curve has an extreme value as a keyframe candidate. As a result of the experiment, it was found that about 80% of the correct keyframes can be automatically extracted.

Keywords: Sign language, video abstraction, keyframes, posture estimation, curve approximation.

1. はじめに

手話を第一言語とする聴覚障がい者の方々の中には、メールや字幕を用いた文字を中心とした情報伝達やコミュニケーションを得意としない人たちが存在する。近年のスマートフォンをはじめとするカメラ付き携帯端末の普及により、ある程度、手話映像を用いたコミュニケーションが可能な環境が整っている。しかし、例えば災害時には通信帯域が圧迫されることが多いため、情報を伝える手話映像などを送る際には、できるだけデータ容量を小さくして伝達することが重要である。

また、手話は一連の動きを見なければ手話の内容を理解することが難しいため、映像を再度確認する際にも時間がかかってしまう。

これらの問題を解決するために、手話映像における一連の動作の内容を損なうことなく要約することが本研究の目的である。そのためには、手話映像中で手話の特徴が強く表れている画像フレーム（キーフレーム）を手話の映像の中から抽出し、キーフレームだけで映像を再構成することにより内容を保持したまま要約する方式を検討している。キーフレームのみで構成された要約映像であっても、通常

の手話映像と同等の内容を伝達できることが先行研究により示唆されており[1-6]、キーフレームで手話の映像を再構築することができればデータ削減に大きな効果があることが期待できる。

キーフレームによる伝達は、データ量以外の面でも、実写映像に基づくため表情などの非手指動作による表現を再現できる点や、手話の種類（地域性や方言など）や言語（日本語、英語など）に依存しない可能性がある点が大きな特徴である。

しかしながら、映像中からキーフレームを自動抽出する手法が確立されておらず、どのような画像をキーフレームとすべきか、また、どのようにしてそのようなキーフレームを自動で抽出するのが大きな課題となっていた。

本研究では、手話においてキーフレームとなりうるポーズでは手の動きが停留する、という仮定に基づき、手話映像中の手話者の手の姿勢を推定し、関節位置の時系列データが極値をとるフレーム位置をキーフレーム候補とする手法を提案する。入江らの研究では左肘関節のX座標・Y座標の時系列データを多項式関数で近似し、それぞれの極値をとるフレームをキーフレーム候補としていたが[10]、本研究では両肘および両手首の4つの関節について、X座標・Y座標の合計8種類の時系列データをそれぞれFourier展開により正弦曲線で近似し、極値をキーフレーム候補と

¹ 崇城大学 Sojo University
² 芝浦工業大学 Shibaura Institute of Technology

する。実験において、適切な Fourier 係数の次数を求め、8 種類の曲線から得られたキーフレーム候補から投票および AND を取ることにより絞り込んだ。

以下、第 2 章でキーフレームの概念を説明し、第 3 章で実験について説明する。第 4 章で実験結果および考察を述べ、第 5 章でまとめる。

2. キーフレームについて

本研究では手話動画において、強く手話の特徴を表している画像（フレーム）をキーフレームと呼んでいる。例えば手話における単語の開始点や終了点、手指の軌道が反転するところのように、その画像だけでも意味が通じるであろう画像を示している。実際にキーフレームだけから再構成した動画を、手話を第一言語とする聴覚障がい者に見てもらっても内容を理解することができることが先行研究により実験的に示されている[1-6]。

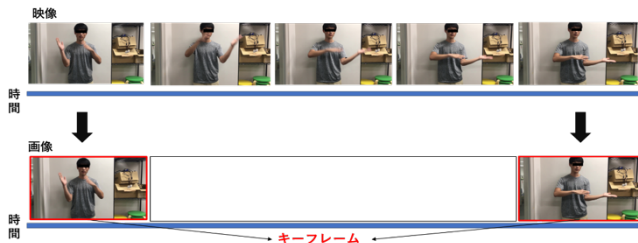


図 1 手話内容「倒れる」 キーフレーム

図 1 はキーフレームの概念を示す図である。ず 1 上部は手話「倒れる」を左から右へ時系列順に表示している（図では 5 枚の画像で表しているが、実際は数十フレームの画像が存在する）。この手話映像から手話の特徴が強く表れている画像フレームを抜き出したものが図の下部にある 2 枚の画像フレームである。この 2 枚の画像で「倒れる」を表現することができれば、数秒の動画を 2 枚の 2 枚の画像で要約することができることになる。これを文章を表す手話映像にも適用できれば、非常に少ないデータ量で、かつ、短い時間で手話映像を再確認できるというメリットがある。

しかし、キーフレームを自動で抽出する手法が先行研究でも確立されておらず、自動抽出は重要な課題である。品田らは、手話において、手指の開始点・終了点や軌道が反転する時点などは、手指の運動が停留するであろうという仮定のもと、手領域の重心の時間変化のグラフが極値をとるところをキーフレーム候補として抽出を試みている[8]。この実験では極値がキーフレームを含むことが示されたが、キーフレームではない候補も多数検出されるという課題があった。また、前述した入江らは 1 つの肘関節位置の時間変化のグラフを多項式近似して極値をキーフレーム候補として検出したが、多項式の次数をどのように設定すれば良いか、という課題があった[10]。

3. 提案手法および実験

本研究で提案する手法の流れは以下の通りである。まず、蓄積された手話映像を対象として OpenPose[7]を用いて両肘部、両手首部の 4 関節の位置の軌跡の抽出を行う。次に、4 関節の X 座標・Y 座標別の計 8 種類の軌跡曲線を次数別に Fourier 級数展開を用いて近似を行い、極値を検出することでキーフレーム候補を抽出する。その後、多数決により次数別に出力した近似曲線の中から、適切な次数の近似曲線の選定を行う。

最後に、AND 法[10]を用いて X 座標と Y 座標別にキーフレーム候補を絞り込み、あらかじめ用意していた実験対象の手話映像の正解キーフレームと比較を行い妥当性を検討する。

3.1 対象とする手話映像

今回実験対象とした手話映像は「私の妻は料理がとてもおいしい」というものである。手話を理解でき、手話で会話ができる人に確認しながら手動でキーフレームを抽出すると、「私（の）」、「妻（は）」「料理（が）」「（とても）おいしい」の 4 語で合計 10 フレームとなった。図 2 に正解のキーフレーム画像を示す。正解のキーフレーム画像の左上の数字は時系列順を示している。



図 2 実験対象の手話映像の正解キーフレーム

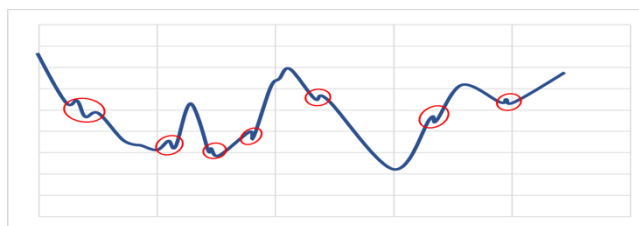


図 3 近似対象の軌跡

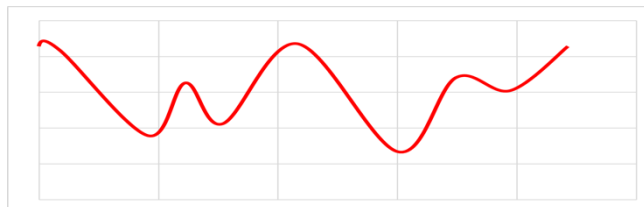


図 4 近似後の軌跡

3.2 提案手法

キーフレームは手話の動作中の手部の座標が時間軸に対して極値（極大値・極小値）をとるようなフレームに包括されていることが先行研究で示されている[8,9].

しかし、実際には図 3 の赤い丸で囲んだ部分のように OpenPose による推定値に小さなノイズが発生してしまい、キーフレーム候補と誤検出されるフレーム数の増加を生じてしまう。そのため、ノイズを除去するために図 4 のように曲線近似を用いて平滑化を行う。

なお、図 3 と図 4 は縦軸が座標値（単位：ピクセル）で横軸を時間（フレーム）としている。

次に近似曲線の最適な次数を求めるため、近似曲線の極値が正解に近いかどうかを点数化処理で評価する。この方法では、あらかじめ用意していた実験対象の手話映像の各フレーム番号に点数を付与しておく。正解キーフレームには 3 点、正解キーフレーム前後の「正解フレームと許容できる許容範囲」のフレーム番号には 2 点、2 点の範囲の前後 1 フレームには 1 点、それ以外には 0 点を配置することで採点を行う。以下の表 1 に多数決法の処理の例を表 2 に表 1 の結果を示す。この採点結果による評価方法を以下、多数決法と呼ぶことにする。

この多数決法では、正解のキーフレームを多く含むように近似曲線の次数を決定するが、これは元の曲線がどのような形をしていると、どれくらいの次数で適切に近似できるのかというようなパターンを学習することができると考えられる。

続いて、多数決法で選定された両肘部と両手首部の 4 関節の近似曲線を X 座標と Y 座標別に分けて AND 法を行う。この AND 法では 8 種類の近似曲線から選定された曲線に対し、（キーフレーム候補となる）極値をとるフレーム番号だけを抜き出す。その後、その候補フレーム番号に対して

表 1 多数決法の処理の例

フレーム番号	極値	点数	採点
1		1	
2		2	
3	○	3	3
4		2	
5	○	2	2
6		2	
7		1	
8	○	0	
9		0	
10		0	

表 2 表 1 の多数決の結果

点数	頻度	採点
0	1	0
1	0	0
2	1	2
3	1	3
	合計	5

一定の閾値を設け、その閾値の範囲内にある他の近似曲線の極値フレーム番号を求め、複数の近似曲線から一定数の候補が抽出されるかどうかで候補とするかどうかを判定する。

表 3 に AND 法の例を、表 4 に表 3 の結果抽出されたキーフレーム候補のフレーム番号を示す。

表 3 閾値 1 で 2 関節以上での AND 法の例

右肘	右手首	左肘	左手首	AND法の結果
6	2	4	10	6
13	6	8	23	6
21	16	7		7
26	25	13		25,26
29	31	18		31
31	40	31		31
37				
44				
極値フレームのフレーム番号				

表 4 表 3 の結果

2関節以上のAND法の結果				
6	7	25	26	31

本研究では AND 法を多数決の後に行うため、多数決を行う際には、他の次数の近似曲線よりも合計の点数が平均以上であり、かつ、2 点と 3 点の数が平均よりも多く、0 点の数が平均よりも少ないような近似曲線を選ぶことにした。同じ条件の場合は次数の少ないほうの曲線を採用する。これにより、適切な次数の近似曲線で AND 法を行うとどのような結果で出るのか検証を試み、表 4 の結果のようにフレーム番号の画像フレームを出力して、正解のキーフレームと比較を行い評価を行なった。

4. 実験の結果

本実験では実験対象の手話映像の「私の妻は料理がとってもおいしい」に対して、「フーリエ級数展開を用いた近似」「多数決法」「AND 法」の 3 つの実験を行った。

図 5 に今回の実験対象の手話映像の「私の妻は料理がとってもおいしい」の右肘の X 座標の曲線（縦軸は座標値、横軸は時間（フレーム））を、図 6 に次数別に近似した曲線を示す。図 6 の各近似曲線グラフの中心にある数字は近似曲線の次数を示している。

表 5 は右肘の X 座標の多数決法を施したものである。表 5 から、合計の点数が平均以上、かつ 2 点・3 点の数が平均以上、かつ 0 点の数が平均以下、であるような近似曲線は次数 12 であるため出力される近似曲線は次数 12 を選定した。このような処理を、両肘部と両手首部の 4 関節を X 座標と Y 座標別に計 8 種類行なった。

図 7 に 4 関節の X 座標に対し、閾値 1、かつ、2 関節以上の条件で行った AND 法によって出力された中から正解キーフレーム、および、正解キーフレームと許容できるフレームのみを示す。図 8 には図 7 と同じ条件で行った Y 座標の AND 法によって出力された中から正解キーフレーム、および正解キーフレームと許容できるフレームのみを示す。

5. 考察

本実験での最終的な結果である AND 法では正解キーフレームと、正解キーフレームと許容できるフレームのみを手動で出力した。X 座標での結果では「私（の）」部分が、Y 座標では「料理（が）」の部分が一部欠落した。また、X 座標で不要な画像フレームが 10 枚、Y 座標での結果では不要な画像フレームは 26 枚出力されてしまった。

しかし、結果としては正解キーフレームと、正解キーフレームと許容できるフレームを 8 割程度出力することができている。残りの 2 割に関しては、手話の動作上で肘や手首が X 座標や Y 座標に対して変動しない部分があることや、AND 法での条件で抜き出すことができなかったことが考えられる。

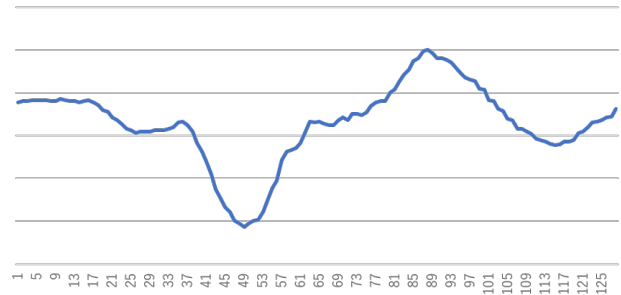


図 5 右肘の X 座標の軌跡

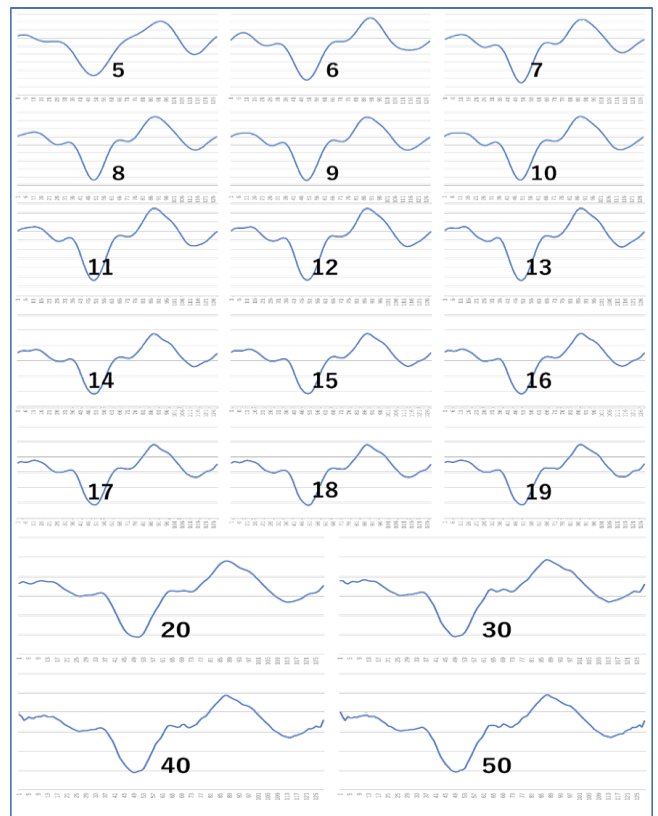


図 6 図 5 の曲線の次数別の近似曲線

したがって、キーフレームでないにも関わらず検出される（誤検出）フレームの問題と、X 座標の結果、Y 座標の結果をどのようにして結合させてお互いに欠落した部分を解決するか、が課題となった。

6. まとめ

本研究では、手話映像の「私の妻は料理がとってもおいしい」を対象として、「フーリエ級数展開を用いた近似」「多数決法」「AND 法」の 3 つの手法を適用し、次数の選定やキーフレーム候補の抽出を行った。

表 5 右肘 X 座標の多数決結果

点数	頻度	合計	点数	頻度	合計	点数	頻度	合計
0	2	0	0	1	0	0	1	0
1	0	0	1	0	0	1	1	1
2	3	6	2	4	8	2	2	4
3	1	3	3	1	3	3	2	6
回数5	計	9	回数6	計	11	回数7	計	11

点数	頻度	合計	点数	頻度	合計	点数	頻度	合計
0	1	0	0	1	0	0	3	0
1	1	1	1	2	2	1	2	2
2	2	4	2	4	8	2	3	6
3	2	6	3	1	3	3	0	0
回数8	計	11	回数9	計	13	回数10	計	8

点数	頻度	合計	点数	頻度	合計	点数	頻度	合計
0	3	0	0	3	0	0	4	0
1	1	1	1	1	1	1	1	1
2	4	8	2	5	10	2	3	6
3	0	0	3	3	9	3	2	6
回数11	計	9	回数12	計	20	回数13	計	13

点数	頻度	合計	点数	頻度	合計	点数	頻度	合計
0	3	0	0	3	0	0	4	0
1	1	1	1	1	1	1	0	0
2	6	12	2	6	12	2	5	10
3	2	6	3	2	6	3	3	9
回数14	計	19	回数15	計	19	回数16	計	19

点数	頻度	合計	点数	頻度	合計	点数	頻度	合計
0	5	0	0	5	0	0	6	0
1	1	1	1	1	1	1	0	0
2	6	12	2	6	12	2	3	6
3	2	6	3	2	6	3	3	9
回数17	計	19	回数18	計	19	回数19	計	15

点数	頻度	合計	点数	頻度	合計
0	8	0	0	10	0
1	1	1	1	2	2
2	4	8	2	4	8
3	3	9	3	4	12
回数20	計	18	回数30	計	22

点数	頻度	合計	点数	頻度	合計
0	15	0	0	21	0
1	2	2	1	1	1
2	5	10	2	6	12
3	5	15	3	5	15
回数40	計	27	回数50	計	28



図 7 X 座標の閾値 1 で 2 関節以上での AND 法の結果

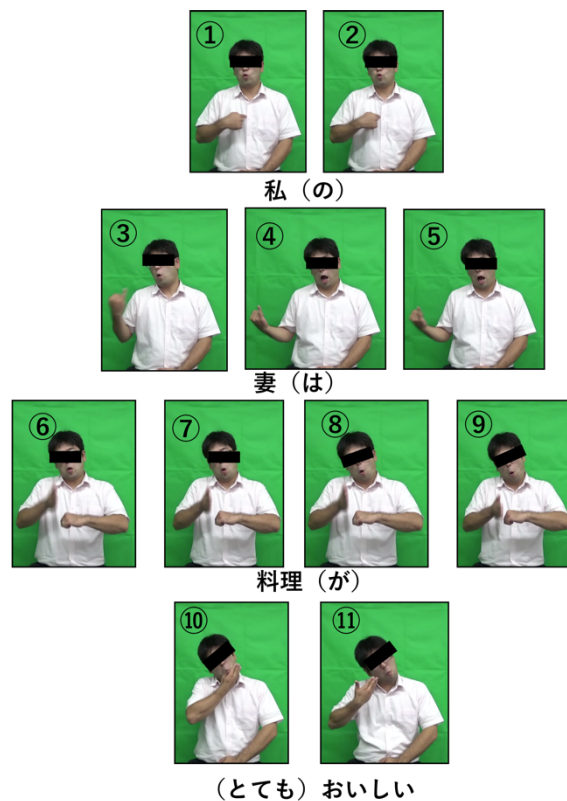


図 8 Y 座標の閾値 1 で 2 関節以上での AND 法の結果

その結果、適切な次数の近似曲線を用いて処理を行えば正解フレームと、正解フレームと許すできる範囲内のフレームを複数枚抽出することができるということを示すことができた。

しかし、本実験では適切な曲線の次数を正解との比較で選定したため、選定した次数が他の手話映像にも適用できるかどうかを確認する必要がある。重要な課題としては、多数決の実験を重ねることで適切な次数の設定に必要な極値の数や、極値の横軸である時間（フレーム）に対する幅や、振幅の大きさ等の相関性を見つけることであると考ええる。この課題を解決することで、AND 法を行う際に不必要なフレームを除外することができるようになり自動化に近づくと考えられる。

謝辞

本研究は科研費（基盤研究（C））19K2032「手話映像の時間的要約方式に関する研究」に基づくものである。また、実験を行うにあたってご協力いただいた NTT クラリティ株式会社の関係者のみなさま、芝浦工業大学 猪岡翔氏に感謝する。

参考文献

- [1] 秋山滉太, 筒口拳, 米村俊一: “キーフレーム通信方式を用いた災害時情報伝達システムの提案”, 信学技報 (福祉工学研究会) (Aug. 2015).
- [2] 秋山滉太, 筒口拳, 米村俊一: “手話のキーフレームに基づく映像圧縮を用いた災害情報伝達システム”, 情報処理学会全国大会 (Mar. 2016).
- [3] 秋山滉太, 筒口拳, 米村俊一: “手話の空間的特徴に基づく映像圧縮を用いた災害情報伝達システム: 20名のろう者によるキーフレーム映像の有効性評”, 信学技報 (ヒューマンコミュニケーション基礎研究会) 116(31), 201-206 (May 2016).
- [4] 秋山滉太, 筒口拳, 米村俊一: “手話の空間的特徴に基づく映像圧縮を用いた災害情報伝達システム ~キーフレーム映像による手話伝達の了解度の検証~, ヒューマンインタフェースシンポジウム 2016 (Sep. 2016).
- [5] 秋山滉太, 筒口拳, 米村俊一: “手話の空間的特徴に基づく映像圧縮を用いた災害情報伝達システム ~無圧縮映像における手話の了解度についての考察~, 第87回福祉情報工学研究会 (Dec. 2016).
- [6] 秋山滉太, 筒口拳, 米村俊一: “手話の空間的特徴に基づく映像圧縮を用いた災害情報伝達システム ~手話映像の繰り返し再生におけるろう者の情報取得戦略 ~”, 第91回福祉情報工学研究会 (Aug. 2017).
- [7] OpenPose LibrarY: <https://github.com/CMU-Perceptual-Computing-Lab/openpose>.
- [8] 品田紗弥花, 筒口拳, 米村俊一: “カラー手袋を用いた手話の空間的特徴抽出方法に関する基礎検討”, ヒューマンコミュニケーション基礎研究会 (May 2017).
- [9] 品田紗弥花, 筒口拳, 米村俊一: “両手手話の利き手に着目した手話キーフレーム抽出方式の検討”, 信学技報, vol. 118, no. 49, HCS2018-28, pp. 197-202 (May 2018).
- [10] 入江健太, 米村俊一, 筒口拳: “関節軌道の多項式近似に基づく手話映像からのキーフレーム抽出”, 画像電子学会 292 回研究会, 2020 年 2 月.