論文 / 著書情報
Article / Book Information

論題	ToFカメラによる3D手話認識
Title	
著者	佐藤 新, 篠田 浩一, 古井 貞熙
Author(s)	Arata Sato, Koichi Shinoda, SADAOKI FURUI
出典	画像の認識・理解シンポジウム（MIRU2010), IS3-44, No. , pp. 1861-1868
Citation	, IS3-44, No. , pp. 1861-1868
発行日 / Pub. date	2010, 7

ToF カメラによる 3D 手話認識

佐藤　新[†]　篠田 浩一[†]　古井 貞熙[†]

† 東京工業大学大学院情報理工学研究科 〒 152-8550 東京都目黒区大岡山 2-12-1
E-mail: †sato@ks.cs.titech.ac.jp,{shinoda,furui}@cs.titech.ac.jp

あらまし　本稿では，Time-of-Flight(ToF) カメラから得られる 3 次元情報を用いて手話を認識する手法を提案する．ToF カメラとは被写体までの距離を測定することができるカメラであり，近年，さまざまな分野への応用が試みられているデバイスである．コンピュータによる手話の自動認識の従来研究としては，普通のビデオカメラである 2D カメラを用いる手法と，磁気や加速度のセンサを用いる手法があるが，2D カメラを用いた手話認識の場合は，外乱の影響を受けやすいという問題があり，一方で，センサを用いた手話認識の場合は，カメラの場合と違って手の形を求めることができず，さらに，手にセンサを取り付けなければならないといった問題がある．これらの問題は，ToF カメラを用いることにより解決することができ，それによって認識率の向上が期待できる．評価実験として，手話を 2D カメラと ToF カメラの両方で同時に収録し，それらのデータからそれぞれ計算した特徴量を隠れマルコフモデルを用いて認識する比較実験を行った．実験の結果，2D カメラを用いた手話認識の正解率は最高で 58.7%だったのに対し，ToF カメラを用いた手話認識の正解率は最高で 91.3%となり，提案手法の有用性を確認した．
キーワード　手話認識，Time-of-Flight カメラ，隠れマルコフモデル

Sign Language Recognition Using Time-of-Flight Camera

Arata SATO[†], Koichi SHINODA[†], and Sadaoki FURUI[†]

† Tokyo Institute of Technology, Graduate School of Information Science and Engineering
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, JAPAN
E-mail: †sato@ks.cs.titech.ac.jp,{shinoda,furui}@cs.titech.ac.jp

Abstract　We propose an automatic sign language recognition system using a Time-of-Flight(ToF) camera. A ToF camera can measure depth of subjects and it has been applied to various tasks. Conventional automatic sign language recognition methods are classified into two categories. One is those using a 2D camera and the other is those using sensors such as magnetic sensors or accelerometers. Those methods using a 2D camera are not robust against noises, while those methods using sensors cannot obtain hand shape parameters. Moreover signers need to mount sensors on their hands. With a ToF camera, the problems in the conventional methods using a 2D camera or sensors can be solved. We compared our method with a method using a 2D camera. While correctness was 58.7% by using a 2D camera, it was 91.3% by using a ToF camera. This result confirmed effectiveness of our method.
Key words　Sign language recognition, Time-of-Flight camera, Hidden Markov Model

1. はじめに

近年，障がい者の社会進出が進んでいる．そのような背景において，障がい者とコミュニケーションをとることは非常に重要である．聴覚障がい者との場合，筆談でコミュニケーションを行うことも可能だが，効率を考えると手話を用いてコミュニケーションができるほうが有用である．しかしながら多くの人は手話に対する知識がないため，コンピュータによる手話の自動認識が求められている．

手話はそれ自体が文字を持たないものの，言語の一種であるということができる [15]．つまり手話を認識する手法としては，同じ言語である音声を認識する手法，すなわち隠れマルコフモデル (HMM) を用いた認識手法を応用することができる．したがって問題は，手話の時系列データからどのような特徴量を抽出するかである．

従来の代表的な手話の特徴量抽出方法としては，通常のビデオカメラである 2D カメラによって得られたデータから特徴量を抽出する方法と，手に取り付けたセンサで計測したデータから特徴量を抽出する方法の二通りが存在する．

2D カメラからの特徴量抽出では，肌の色を検出して手の位置と形を推定する研究 [1] や，色だけでなく輪郭も用いて手の位置と形を推定する研究 [2] が行われてい

るが，肌の色には個人差があること，また照明の影響を大いに受けること，背景に似たような色のものがあるとそれを誤認識してしまうことなどからロバスト性に問題が生じてしまう．また，カメラに対して前後に動くような手の動きを検知することができないという問題もある．この問題に対してはステレオカメラを用いて対処することも可能だが，ステレオカメラ画像からの距離計算は，ステレオマッチングなどの計算量の多い処理が必要になり，リアルタイム性の確保がしばしば困難である．

一方で，センサからの特徴量抽出では，磁気センサから得られた位置のデータから特徴量を抽出する研究 [4] [5] や，加速度センサから得られた加速度のデータから特徴量を抽出する研究 [7] が行われてきた．これらの手法では外乱の影響を受けにくい点，手の前後の動きを観測できる点では 2D カメラによる手話認識より優秀であるが，入力が画像でないため手の形を求めることができない．また，手にセンサを取り付ける必要があり，それが手話話者の負担となる可能性がある．

そこで，本稿では ToF(Time-of-Flight) カメラを用いた手話認識を提案する．ToF カメラとは被写体までの距離を赤外線を用いて測定することができるカメラであり，近年，量産化が進んだことによって比較的安価で手に入れられるようになったデバイスである．その結果，現在では自動車における歩行者のモニタリングや，巨大スクリーンのインターフェースなど，さまざまな分野への応用が試みられている [17]．この ToF カメラを用いれば，磁気センサや加速度センサによる手話認識同様に手の位置や動きを三次元で測定することができ，また，2D カメラの問題点であった外乱の影響を受けることなく，手の形を測定することが可能である．

ToF カメラを用いたパターン認識の研究としては，例えば，指文字認識 [13] がある．この研究では，認識対象を指文字としているために片手しか認識を行っていない．指文字に加え，簡単なジェスチャー認識を行っている研究 [12] もあるが，こちらも手が片手であるか両手がくっついた状態であることを前提としている．また，ToF カメラから得られた距離画像に対して手のモデルフィッティングを行っている研究 [8] もあるが，正確なモデルフィッティングを行っているため計算処理が非常に重い．それゆえにリアルタイム性が確保できず，実際のジェスチャー認識には至っていない．

本稿の手法では，ToF カメラで撮影された距離画像から手領域を抽出し，その領域から計算した手の三次元座標と，距離画像から計算した手の形の特徴量を用いて，HMM で手話認識を行う．また，従来研究とは異なり，両手を使った手話も認識の対象とする．ToF カメラを用いることで 2D カメラやセンサによる手話認識の問題点を克服し，さらなる手話の認識率の向上を目指す．

図 1　SR4000 外観

2. ToF カメラ

ToF カメラとは，Time-of-Flight(飛行時間) 計測原理に基づいて距離を計測するカメラである．

2.1 原　　理

ToF カメラは，カメラ自体に付属している発光源から視野内の対象物へ光を放出し，センサーへ帰還するまでの光の飛行時間を計測することによって距離を計算する．

2.2 SR4000

今回使用した ToF カメラである SR4000 について説明する．

2.2.1 概　　要

Time-of-Flight による距離計測を実現するために，SR4000 は，自身の LED 光源を変調させ，CCD/CMOS イメージセンサは各画素ごとに反射してきた変調信号の位相を計測する．各画素での距離は，その変調信号の周期の分数として定義される．距離 D と周期は次式によって対応する．

$$D = \frac{c}{2f} \tag{1}$$

ただし，c は光速，f は変調周波数を表す．SR4000 の初期設定上の変調周波数は 30MHz なので，5.00m までの距離計測が可能である．その他にも，反射強度の計測をすることができ，それによってグレースケール画像を得ることも可能である．

SR4000 の主な仕様 [16] を表 1 に示す．ToF カメラの解像度は 2D カメラよりも劣ってはいるものの，Sperling らによれば，フレームレートが 15fps の 24×16 ピクセルの動画であっても，人は 85%の精度で理解することができる [11] とあることから，解像度による認識性能の差は少ないと考えられる．また，解像度以外の点に関しては，2D カメラの場合と比較しても手話認識に関して著しく劣っている点はないといえる．

2.2.2 問題点とその解決策

ToF カメラの問題点とその解決策について説明する．

仕様		
製造元	MESA Imaging AG	
画素配置	176 (h) x 144 (v)	QCIF
視野角	43.6 ° x 34.6 °	
画素サイズ	40 μ m	水平垂直ともに
角度分解能	0.23 °	中央の画素にて
発光波長	850nm	波長中央値
変調周波数	30 MHz	初期設定値
動作範囲	0.3 ～ 5.0 (m)	標準設定にて
距離精度	+/- 1 cm	z-方向, 1 画素単位
フレームレート	最高 54 FPS	カメラ設定に依存

表 1　SR4000 仕様

距離画像の座標変換

ToF カメラで得られる距離画像は，被写体までの直線距離を表すものであり，したがって円形にゆがんでしまうという問題がある．しかし，SR4000 では直線距離を 3 次元座標に自動でキャリブレーションを行うため，物体の正しい位置座標を得ることができる．

ノイズ

ToF カメラで得られる距離画像には多くのノイズが現れるが，SR4000 ではそれぞれの画素で近傍の反射強度や距離情報を統合することによってノイズの影響を低減した距離画像を取得することができる．

測定範囲外の値

ToF カメラは，光の位相のずれから距離を計算するため，測定範囲外の距離を誤って計測してしまう問題がある．例えば最大測定距離が 5m の場合，ToF カメラは 5m 先の位置にある物体も 10m 先の位置にある物体も計測結果は同じ 5m となってしまう．SR4000 でもこの問題は生じるが，光の反射強度を閾値としてフィルタリングすることによって，最大測定距離内の正しい値だけを得ることができる．

3.　手話認識システム

ToF カメラによる手話認識システムは以下の流れで処理を行う．まず最初に ToF カメラによって得られる距離画像から手領域を抽出し，次にその領域を掌や腕にクラスタリングする．クラスタリングされた各々の領域から特徴量を計算し，最後にそれらの特徴量を入力として隠れマルコフモデルで認識を行う．上述のシステムの各処理の詳細を以下で述べる．

3.1　手領域抽出

ToF カメラから得られた距離画像からの手領域の抽出は，奥行き方向の座標でフィルタリングすることで行う．すなわち，ある奥行き座標を閾値として，その閾値よりも手前に写っている物体が手の領域であると考えることにする．なお，今回閾値は手動で設定した．

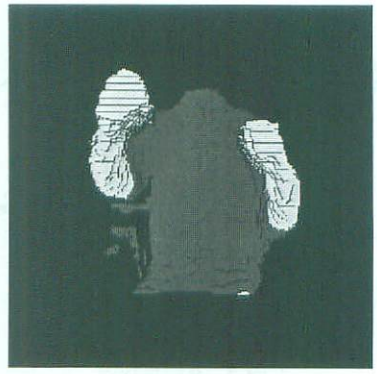

図 2　奥行き座標でのフィルタリングによる手腕領域の抽出

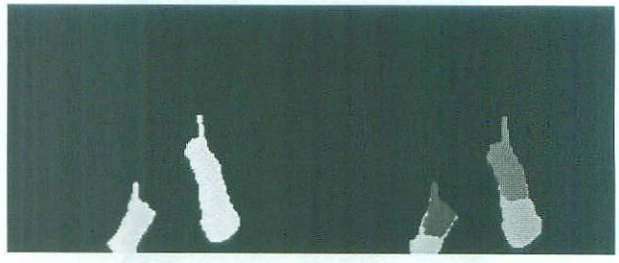

図 3　抽出した手領域（左）をクラスタリングした例（右）

3.2　クラスタリング

次に，抽出した部分を右掌，右腕，左掌，左腕の 4 つにクラスタリングする．次のような場合分けによってクラスタリング方法を変える．

片手のみが存在する場合

抽出した領域の重心が中央より左にあれば，その領域は右手であると判断する．逆に右にあれば，その領域は左手であると判断する．次に領域を k-means 法を用いて二つに分割し，手前や上側にあるクラスタを掌，そうでないほうを腕と判断する．

両手が離れて存在する場合

二つの領域の重心を比較して，左側にある領域を右手であると判断し，右側にある領域を左手であると判断する．それぞれの領域を k-means 法を用いて二つに分割し，手前や上側にあるクラスタを掌，そうでないほうを腕と判断する．

両手がくっついて存在する場合

抽出した領域を k-means 法で三つに分割し，左側にあるクラスタを右腕，右側にあるクラスタを左腕であると判断する．中央のクラスタは左右どちらの掌にも属することにする．

ただし，上記によるクラスタリングのみの場合，誤認識が発生しやすい．例えば，左手を右に動かしていった場合，この方法では途中からその手は右手であると判断されてしまう．そのような誤認識は，前回状態を参照することによって回避することができる．よって，状態遷

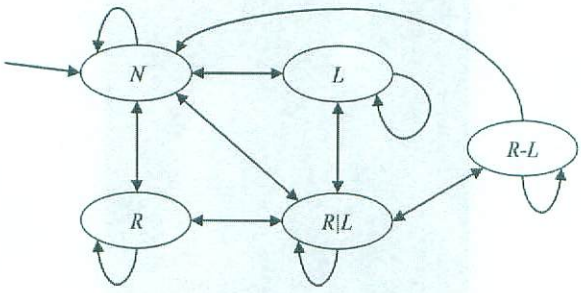

図 4　状態遷移図

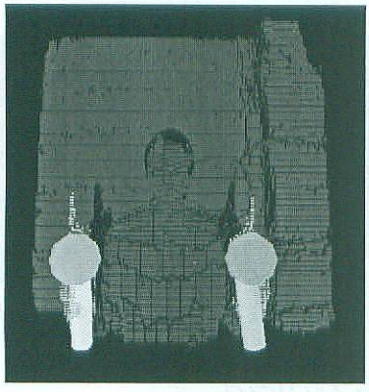

図 5　PCA による腕の方向推定：PCA による腕方向の計算結果を円柱で可視化した．掌は球で可視化している．

移を導入する．

今回の場合，次の五つの状態を定義する．

- N : 右手も左手も存在しない．
- R : 右手のみが存在する．
- L : 左手のみが存在する．
- $R|L$: 両手が離れて存在する．
- R-L : 両手がくっついた状態で存在する．

可能な状態遷移は図 4 のようにする．このように状態遷移を制限することによって，先程の例のような誤認識を回避することができる．

3.3　特徴量抽出

次に，3.2 によって決定した各々の領域から特徴量を計算する．

まずは，各々の領域の重心座標を対応する掌や腕の 3 次元位置座標とし，その時間差分を掌や腕の速度とする．また，領域は 3 次元座標をもつ点の集合で与えられているので，腕領域に対して PCA（主成分分析）を適用して 1 次元に次元圧縮し，このとき得られた固有ベクトルを腕の向きとする．さらに，掌や腕の 3 次元座標や腕の向きの特徴量はカルマンフィルタを用いて補正を行う．

掌領域からは，手の形に関する特徴量を計算する．まずは掌領域画像から掌を含む矩形を検出し，以下のものを特徴量として出力する．

- 矩形内における掌領域の割合

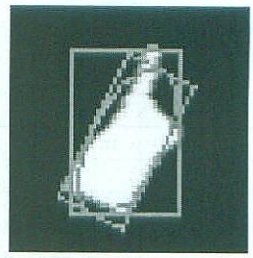

図 6　掌領域からの特徴量抽出：掌を含む矩形や，楕円近似の結果を表す

- 矩形の 8 方向の各部分における掌領域の割合
- 掌領域に対する楕円近似時の長軸と短軸の比
- 掌領域に対する楕円近似時の軸の角度

求めた特徴量をまとめると，以下のようになる．

- 手の有無
- 腕の重心座標（3 次元）
- 腕の重心座標の変化量（3 次元）
- 腕の方向ベクトル（3 次元）
- 腕の方向ベクトルの変化量（3 次元）
- 掌の重心座標（3 次元）
- 掌の重心座標の変化量（3 次元）
- 掌領域を含む矩形内における掌領域の割合
- 掌領域を含む矩形の 8 方向の各部分における掌領域の割合（8 次元）
- 掌領域に対する楕円近似時の長軸と短軸の比
- 掌領域に対する楕円近似時の軸の角度

よって，片手につき 30 次元，両手で 60 次元の特徴量が得られたことになる．

3.4　隠れマルコフモデル (HMM) による認識

最後に各手話単語を隠れマルコフモデル (HMM) でモデル化し，3.3 で求めた特徴量を入力として手話認識を行う．HMM の各状態の出力分布は単一ガウス分布とし，HMM の構造は状態数 15 の Left-to-Right 型 HMM と状態数 8 の 1 状態スキップありの Left-to-Right 型 HMM の 2 通りで認識を行った．学習と認識には HTK-3.4 [18] を使用した．

4.　評価実験

今回の実験では，2D カメラと ToF カメラの両方で同時に手話話者を撮影し，そのデータからそれぞれ特徴量を計算し，HMM を用いて学習，認識を行う．

4.1　実験条件

4.1.1　実験データ

今回実験用に撮影したデータは以下のようなものである．まず，被験者には 2D カメラが明らかに不利にならないように黒い服を着てもらった上で，カメラの前の椅子に座ってもらう．最初は両手をひざの上においた状

会う	家	行く	美しい	今日	来る
さようなら	～する	どちら	場所	前	休み

表 2　収録した手話単語一覧

図 7　手話の例：会う

図 8　手話の例：どちら

図 9　手話の例：行く

図 10　手話の例：場所

態から撮影を開始し，手話を行った後に両手をひざの上に戻したら撮影を終了する．ただし，撮影した動画のフレームレートは，ToF カメラは 10fps であるのに対し 2D カメラは 60fps である．被験者は普段手話を使わない右利きの健聴者 5 名で，収録を行う手話の動作は，事前に [14] 付属の DVD の動画で確認している．収録した手話は，表 2 に示す 12 単語であり，1 単語につき，学習用データを 10 個，テスト用データを 5 個収録した．したがって，最終的な学習データ数は 600 個，テストデータ数は 300 個となった．

収録した手話は，ある程度動作が似ている手話や前後方向への動作があるものを選んだ．例えば「会う」「どちら」の二つの手話では，手の形が同じであるが動作が異なっている．逆に「行く」「場所」「さようなら」といった手話では，手の位置や動きが似ているが手の形や向きが異なっている．また「来る」「～する」「前」といった手話は前後の動きが比較的大きい手話である．

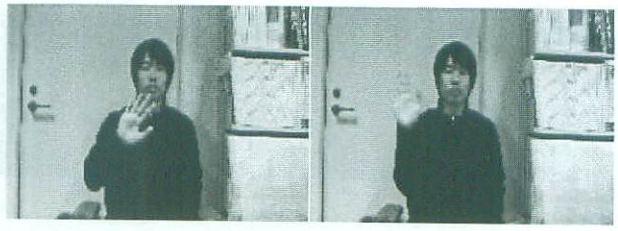

図 11　手話の例：さようなら

図 12　手話の例：来る

図 13　手話の例：～する

図 14　手話の例：前

4.1.2　2D カメラの特徴量抽出

ここでは，比較実験のために行う 2D カメラの特徴量抽出方法について簡単に述べる．

手領域の抽出

手領域は画像から肌色の部分を抽出することで行う．具体的には，画像を RGB 形式から HSV 形式に変換し H の値に対して条件を設け，それを満たす領域を手領域であると判断する．ただし，これだけでは顔も抽出されてしまうため，Haar-like 特徴を用いた顔認識により顔領域は除外する．その後の処理は ToF カメラのときとほぼ同様であるが，2D カメラからの画像では，腕部分の推定が困難なため，腕部分に関する計算は行っていない．

特徴量の計算

手の位置や動きの特徴量には，先ほど求めた手領域か

図 15　2D カメラ画像からの手領域の抽出：入力画像（左）から肌色領域を抽出し（中），顔認識によって顔部分を取り除く（右）．右図の円は右手左手それぞれの領域の重心を表している．

ら計算した重心座標と，その重心座標の変化量を用いる．これらは毎フレームごとにカルマンフィルタを用いて推定を行う．手形の特徴量に関しては，手領域に対する楕円近似時の長軸と短軸の比と手領域に対する楕円近似時の軸の角度のみを用いる．これは，動いている手が画像中でぶれてしまうことによりうまく手領域を抽出できない問題を画像全体にぼかしを入れることで解決しているため，それによって詳細な手形を求めることができないからである．

求めた特徴量をまとめると以下のようになる．

- 手の有無
- 掌の重心座標（2 次元）
- 掌の重心座標の変化量（2 次元）
- 掌領域に対する楕円近似時の長軸と短軸の比
- 掌領域に対する楕円近似時の軸の角度

よって，片手につき 7 次元，両手で 14 次元の特徴量が得られたことになる．ToF カメラの場合と同様に，各手話単語を HMM でモデル化し，これらの特徴量により HMM の学習と認識を行う．

4.1.3　評価基準

本実験では，一つのデータには一つの手話単語が対応するということを既知として扱っているので，各モデルの認識性能の評価には正解率 (Correct) を用いた．これは次式によって求めることができる．

$$\mathrm{Correct} = \frac{N-S}{N} \times 100 \quad [\%] \qquad (2)$$

ここで N は認識する手話データの総数，S は置換誤りの個数である．

4.2　実験結果

認識モデルを状態数 15 の Left-to-Right 型 HMM のときの実験結果を表 3,4 に示す．表 3 は 2D カメラからの特徴量で認識したときのコンフュージョンマトリックスであり，表 4 は ToF カメラからの特徴量で認識したときのコンフュージョンマトリックスである．2D カメラのときの正解率は 58.7%，ToF カメラのときの正解率は 85.3%となった．

また，認識モデルを状態数 8 の 1 状態スキップありの Left-to-Right 型 HMM のときの実験結果を表 5,6 に示す．表 5 は 2D カメラからの特徴量で認識したときのコンフュージョンマトリックスであり，表 6 は ToF カメラからの特徴量で認識したときのコンフュージョンマトリックスである．2D カメラのときの正解率は 55.0%，ToF カメラのときの正解率は 91.3%となった．

	会う	家	行く	美しい	今日	来る	さようなら	～する	どちら	場所	前	休み
会う	**14**	4		1	1			1	2		2	
家		**22**							2		1	
行く	1		**10**	1		1	6	2	1	2	1	
美しい	2			**17**	1			1	1	2		1
今日	1			2	**19**				1	1		1
来る	1		3			**10**	5		3	2	1	
さようなら			8	1		2	**10**		3	1		
～する		1		2	2			**19**			1	
どちら	1								**24**			
場所			11	1	1	1	3	2	1	**4**		1
前	2		1	1		5	7		2	1	**6**	
休み					1			1		2		**21**

表 3　コンフュージョンマトリックス（2D カメラ-状態数 15 の Left-to-Right 型 HMM のとき）

	会う	家	行く	美しい	今日	来る	さようなら	～する	どちら	場所	前	休み
会う	**25**											
家		**25**										
行く			**18**							3	4	
美しい				**25**								
今日					**23**			2				
来る						**22**	3					
さようなら							**25**					
～する	2							**23**				
どちら	3								**22**			
場所			11			6	2			**3**	3	
前							2				**23**	
休み					1			2				**22**

表 4　コンフュージョンマトリックス（ToF カメラ-状態数 15 の Left-to-Right 型 HMM のとき）

まず，正解率を比較すると，状態数 15 の Left-to-Right 型 HMM の場合，状態数 8 の 1 状態スキップありの Left-to-Right 型 HMM の場合の両方で，ToF カメラを用いた手法のほうが高い正解率を示している．つまり，ToF カメラのほうが手話認識に適していると言うことができる．

さらに詳しく見てみると，2D カメラの場合は「行く」「さようなら」「場所」といった動きが似通っている手話を正しく認識できていないことがわかる．これは，微妙に変化する奥行きを特徴量として使っていないこと，手の形の特徴量を十分に用いていないことが原因であると考えられる．また「来る」「前」といった前後に手を動かす手話の正解率もあまり高くない．これも奥行きを特徴量として使っていないことに起因するものであると考え

	会う	家	行く	美しい	今日	来る	さようなら	～する	どちら	場所	前	休み
会う	**12**	6				3			3		1	
家		**21**				3			1			
行く			**15**	2			2		1	3	2	
美しい	3			**17**				1		1	1	2
今日	2			1	**18**		1			1		2
来る			1			**10**	6			6	2	
さようなら			5			3	**10**			6	1	
～する	3	2		4	3			**10**	2		1	
どちら	3	2							**19**		1	
場所			5				7	5	1	**7**		
前			1			3	1	2	1	2	**14**	1
休み	1		2	8	1			1				**12**

表 5　コンフュージョンマトリックス（2D カメラ-状態数 8 の 1 状態スキップあり Left-to-Right 型 HMM のとき）

	会う	家	行く	美しい	今日	来る	さようなら	～する	どちら	場所	前	休み
会う	**23**	2										
家		**25**										
行く			**22**								3	
美しい				**25**								
今日				1	**20**			2				2
来る						**22**	2				1	
さようなら							**24**			1		
～する	2							**23**				
どちら	1			1					**23**			
場所			1			1				**23**	1	
前							3				**21**	
休み								2				**23**

表 6　コンフュージョンマトリックス（ToF カメラ-状態数 8 の 1 状態スキップあり Left-to-Right 型 HMM のとき）

られる．

ToF カメラの場合は 2D カメラに比べるとかなり正解率が高い．しかし，一回目の実験では「場所」という手話だけ著しく正解率が低いことがわかる．「場所」という手話は，右掌を曲げて下に向けた状態で，それを上から下に下ろすという動作であるが，実際の撮影したデータを見てみると，人によっては手を上に上げる際に一度後方に手を移動させてから前に出すという動きをしていた．これにより，手が閾値外へ移動してしまって手の位置を推定できなくなってしまったことが正解率が低い原因であると考えられる．二回目の実験では状態をスキップできる HMM をモデルとして用いたため，手の位置が推定できない状態をスキップすることができるようになり，認識率が大幅に改善されたと考えられる．

5. まとめと今後の課題

本稿では，ToF カメラを用いた手話認識システムを提案し，通常のビデオカメラである 2D カメラを用いた場合との認識性能の比較実験を行った．実験の結果，2D カメラを用いた場合の認識正解率は最高で 58.7%だったのに対し，提案手法である ToF カメラを用いた場合の認識正解率は最高で 91.3%となり，提案手法の有用性を確認することができた．

今後の課題としては，以下のような事が考えられる．

- 座標の正規化：本稿では，ToF カメラから得られた位置座標をそのまま使用しているが，手を動かす大きさは，同じ手話でも各個人で異なっている．そこで，これらの座標を正規化することによって認識率の向上が期待できる．
- 閾値の動的変更：本稿では，手領域の抽出を手動で設定した閾値によって行っているが，顔認識等により手話話者の位置を推定することで，閾値を自動的に決定することができると考えられる．
- 体より後方の手の認識：本稿の手法では，体より後方に手が行ってしまった場合に対応ができない．1 これらの問題には，顔や体の部分を認識して取り除くことによって対応することができると考えられる．
- 手の交差時における正しい認識：本稿の手法では，手の交差時に正しい位置の推定が困難となってしまう．これは，前方の腕が後方の腕を隠してしまうことによりクラスタリングが正しく行えないことが原因である．したがって，このような場合にも対応できる手のクラスタリング方法を考える必要があると思われる．
- 顔認識の導入：手話において，顔の表情というものは，手話認識に有効な要素の一つである．よって顔の表情についての特徴量を追加することによって，認識率の向上が期待できる．
- マルチストリーム HMM 等の導入：本稿では，手の位置や動き，形の特徴量を単純に Feature Fusion 法によって結合しているが，これらの特徴量は別々に分けて扱ったほうが良いということがわかっている [3]．したがって，マルチストリーム HMM 等を用いてこれらの特徴量を分けて学習，認識を行うことによって，認識率の向上が期待できる．
- 大語彙認識：本稿では各手話単語を HMM でモデル化して認識を行ったが，大語彙認識を行う場合，単語数だけ HMM が必要になり，認識率が低下してしまう恐れがある．そのため，基本動作を HMM でモデル化することで大語彙認識に対応する手法 [6] を検討する必要があると考えられる．
- 連続単語認識：本稿では孤立単語認識を行ったが，より実用的なシステムを構築するには連続単語認識ができる必要がある．

謝　辞

本研究は科学研究費補助金基盤研究（B）20300063 の援助を受けた。

文　献

[1] T. Starner, J. Weaver, A. Pentland, "Real-time American Sign Language recognition using desk and wearable computer based video," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol.20, No.12, 1998.

[2] O. Diamanti, P. Maragos, "Geodesic active regions for segmentation and tracking of human gestures in sign language videos," *ICIP*, pp. 1096-1099, 2008.

[3] S. Theodorakis, A. Katsamanis, P. Maragos, "Product-HMMs for automatic sign language recognition," *In Proc. ICASSP*, pp. 1601-1604, 2009.

[4] 前畠 大, 西田 昌史, 堀内 靖雄, 市川 熹, "位置と動きの動作要素に基づく手話認識に関する検討," *In Proc. WISS*, pp. 129-130, 2007.

[5] 前畠 大, 西田 昌史, 堀内 靖雄, 黒岩 眞吾, "手の位置と動きに着目した HMM による手話単語の認識," 電子情報通信学会技術報告, PRMU2008-20,pp. 7-12, 2008.

[6] 北村 正, 豊倉 行崇, "日本手話自動認識のための基本動作抽出," 電気通信普及財団研究調査報告書, No.21, pp. 485-491, 2006.

[7] P. Yin, T. Starner, H. Hamilton, I. Essa, J. M. Rehg, "Learning the basic units in American Sign Language using discriminative segmental feature selection," *In Proc. ICASPP*, pp. 4757-4760, 2009.

[8] P. Breuer, C. Eckes, S. Müller, "Hand gesture recognition with a novel IR Time-of-Flight range camera - a pilot study," *In Proc. MIRAGE*, pp. 247-269, 2007.

[9] E.Kollorz, J.Hornegger, A.Barke, "Gesture recognition with a time-of-flight camera," *DAGM*, vol.5, No.3-4, pp. 334-343, 2008.

[10] O. Aran, I. Ari, L. Akarun, B. Sankur, "SignTutor:An interactive system for sign language tutoring," *IEEE MultiMedia*, vol.16, pp. 81-93, 2009.

[11] G. Spering, M. Landy, Y. Cohen, M. Pavel, "Intelligible Encoding of ASL Image Sequences at Extremely Low Information Rates," *Computer Vision, Graphics, and Image Processing*, vol.31, No.3, pp. 335-391, 1985.

[12] X. Liu, K. Fujimura, "Hand Gesture Recognition using Depth Data," *the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 529-534, 2004.

[13] S. Malassiotis, M. G. Strintzis, "Real-time hand posture recognition using range data," *Image and Vision Computing*, 26-7, pp. 1027-1037, 2008.

[14] 谷 千春（監修）, DVD で覚える手話辞典, 池田書店.

[15] 神田 和幸, 基礎から学ぶ手話学, 福村出版.

[16] SR4000 User Manual, `http://www.mesa-imaging.ch/customer/Customer_CD/SR4000_Manual.pdf`.

[17] "SwissRanger SR3000 and First Experiences based on Miniaturized 3D-TOF Cameras," `http://www.mesa-imaging.ch/pdf/Application_SR3000_v1_1.pdf`.

[18] HMM Tool Kit (HTK), `http://htk.eng.cam.ac.uk/`.