

距離動画像列からの特定ジェスチャのスポッティング

Finger Spelling Spotting from depth image sequence

田中 翔平† 加藤 伸子‡ 岡崎 彰夫‡ 福井 和広†
 Shohei Tanaka† Nobuko Kato‡ Akio Okazaki‡ Kazuhiro Fukui†

概要

本稿では動画像列からの特定指文字のスポッティング技術の実現可能性を検証する。手形状の距離動画像列にカーネル直交相互部分空間法を適用し、長さの異なる複数の指文字列を扱うために、複数の時間尺度で識別を行う。提案法の有効性を確認するため、指文字を模擬したジェスチャを用いた簡易実験を行った。その結果、動作長の異なる4つの連続疑似指文字を安定的にスポッティングできることを確認した。

1 まえがき

手話とは、聴覚障害者との重要なコミュニケーション手段の一つであり、手や指、腕によって行われる視覚言語の一種である。手話は、それ独自の語彙と文法体系を持つ一つの言語であるが、全てのコミュニケーションを手話だけで完結することは難しい。それを補うように指文字が用いられている。

個々の指文字は、日本語50音(図1)やアルファベットに対応し、人名・専門用語などの固有名詞を表現する際に用いられる。指文字は主に利き手の形で表現されるが、複雑な手形状を有していることに加え、実際に用いられる際には手形状が連続的に変化するため、手話通訳者のような手話熟達者であっても安定して認識することは極めて難しい。

従来の指文字を認識する研究では、データグローブ等の接触型センサを用いる手法[1]やRGBカメラ等の非接触型センサを用いる手法[2]がある。しかし、これらの手法は手指動作の阻害による不自然さや照明変動によるアピアランスの変化に対する不安定性などの問題がある。一方で、距離画像を用いた手法はこれらの問題にある程度対応することが可能である[3]。

指文字認識の分野における多くの先行研究では、指文字を変化のない単一の手形状として取り扱うことが多い。しかし、実際に指文字を使用する際は、手形状が連続的に変化していく。従って、1文字ずつ指文字を識別する手法では、単語の識別は困難である。以上のような背景から本稿では、特定の連続指文字の検出・識別に向けた、距離動画像からの特定ジェスチャのスポッティングアルゴリズムについて述べる。



図1: 指文字50音

2 距離動画像列を用いた手形状識別

スポッティングを行う際に重要である識別手法について説明を行う。識別には、動画像列の識別を効率良く行えるカーネル直交相互部分空間法(KOMSM)[4]を適用した。本方法は指文字のような複雑な形状を有する3次元物体を高精度に識別できることが示されている[3][5]。KOMSMは、相互部分空間法(MSM)[6]の拡張の一つであり、データ分布を非線形な高次元空間で白色化することで、複雑な分布構造に対応したものである。以下に相互部分空間法における部分空間の生成方法と類似度の計算方法を述べる。

2.1 相互部分空間法(MSM)

MSMはクラスごとのデータ分布を線形部分空間として近似し、それらの辞書部分空間と入力部分空間同士のなす正準角を計算することで識別を行う。

具体的には、各クラスに属する学習データ群をそれぞれ主成分分析(PCA)によって、データ分布を近似する次元 s の辞書部分空間 Q を生成する。識別時は、入力である複数のデータベクトルから次元 t の入力部分空間 P を生成し、辞書部分空間と入力部分空間の間の正準角を類似度として用いる。 $m = \min(t, s)$ としたとき類似度 S は以下の式によって定義さ

† 筑波大学大学院 システム情報工学研究科

Graduate School of System and Information Engineering, University of Tsukuba

‡ 筑波技術大学 産業技術学部

Faculty of Industrial Technology, Tsukuba University of Technology

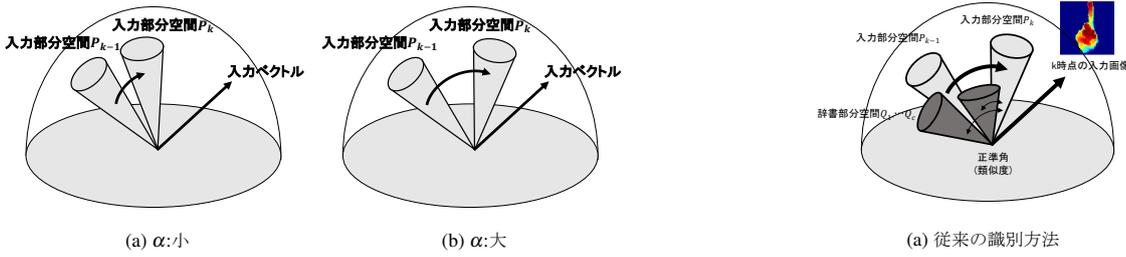


図 2: 部分空間の更新の概念図

れる。

$$\cos^2 \theta_i = \max_{\mathbf{u}_i \in P, \mathbf{v}_j \in Q} \frac{(\mathbf{u}_i \cdot \mathbf{v}_j)}{\|\mathbf{u}_i\| \|\mathbf{v}_j\|}, \quad (1)$$

$$S = \frac{1}{m} \sum_{i=1}^m \cos^2 \theta_i. \quad (2)$$

MSM の拡張である KOMSM の説明については、文献 [4] を参照されたい。

2.2 入力部分空間の更新

本研究では、MSM ベースの手法を用いるため、手領域画像が新たに入力として与えられる毎に入力部分空間を更新する必要がある。入力部分空間の次元は特徴空間の次元に比べて非常に小さいので、同時反復法 [7] を用いて入力部分空間の基底を高速に更新する。処理の流れを以下に記述する。

新たに k 時点で得られた入力画像 \mathbf{x}_k に対して入力パターン分布の自己相関行列 \mathbf{R}_{k-1} を次式により更新する。

$$\mathbf{R}_k = (1.0 - \alpha)\mathbf{R}_{k-1} + \alpha\mathbf{x}_k\mathbf{x}_k^T. \quad (3)$$

ここで \mathbf{R}_k が更新された自己相関行列、係数 $\alpha (0 < \alpha \leq 1.0)$ は、新しい入力画像に対する重み係数である。この値が高いほど、入力部分空間を新たに入力されたベクトルに、より近くなるように回転させる (図 2)。 $k-1$ 時点における入力部分空間の基底を列に持つ \mathbf{E}_{k-1} と更新された自己相関行列 \mathbf{R}_k から次式により行列 \mathbf{E}_k^* を計算する。 \mathbf{E}_k^* は更新された k 時点の自己相関行列の固有ベクトルを列として持つ。

$$\mathbf{E}_k^* = \mathbf{R}_k \mathbf{E}_{k-1}. \quad (4)$$

ただし更新された固有ベクトル \mathbf{E}_k^* は直交条件が崩れているため、Gram-schmidt の直交化変換 GS を用いて、最終的に \mathbf{E}_k の各列を更新された基底とする。

$$\mathbf{E}_k = GS(\mathbf{E}_k^*). \quad (5)$$

この一連の処理によって得られた \mathbf{E}_k を、 k 時点の入力部分空間を張る基底として扱う。

3 提案手法:入力部分空間の並列更新

識別対象が異なる動作長を持つ場合、入力部分空間を完全にカバーするフレーム数が一定ではない。本稿では、部分空

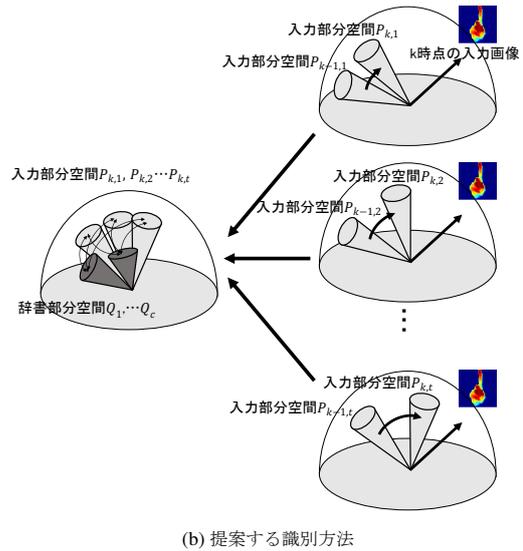


図 3: 更新された入力部分空間の識別の概念図

間の更新パラメータ α を複数用いることで、異なる動作長をもつ対象の識別を行う。具体的には、図 3b のように複数の更新係数で並列に更新される入力部分空間を持つことを提案する。例えば、辞書クラス数が c 、並列に更新される入力部分空間の数が t であるとき、1つの辞書に対して、 t 個の類似度が計算され、結果として、 $c \times t$ 個の類似度が得られることになる。これらの類似度の扱い方は簡単には、最大値・平均・多数決などが考えられる。例えば、最大値をとるというのは、ある辞書に対して入力部分空間 t 個の類似度が得られたとき、その t 個の内最大の類似度をその辞書に対する類似度とする。これをそれぞれの辞書に対して行い、識別を実現する。

本提案手法は、図 3b のように入力部分空間をそれぞれの係数 α について独立に生成することが可能であるので、並列計算の観点からも望ましい性質を持つ。

4 実験

提案法の有効性を検証するための実験について述べる。実験としては、登録データ (学習データ) として 4 種類の連続疑似指文字を学習し、テストデータから検出・識別を行った。撮影には SoftKinetic の TOF 形式の距離センサ DS325 を利用し、切り出された手領域画像をラスタ走査によりベクトル化を行った。このベクトルを特徴量として、KOMSM による識別を行った。以下では、実験に用いた画像列、実験結果と考

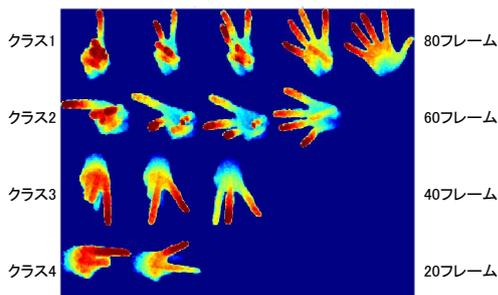


図4: 連続疑似指文字の概要:それぞれの行がクラスに対応しており、立てる指の本数を変えて連続疑似指文字の長さを変える。クラスごとのフレーム数はクラス1から順に80フレーム、60フレーム、40フレーム、20フレームに設定した。

察を示す。

4.1 学習データ

スポッティング対象として、いくつかのジェスチャを連続疑似指文字として定義する。手の方向、上下左右をそれぞれのクラスとし、指を人差し指、中指、薬指、小指、親指の順番に立てることで図4のようにそれぞれのクラスを定義する。図4はそれぞれの行がクラスに対応しており、それぞれのクラスについて240枚、4クラス合計 $240 \times 4 = 960$ 枚のデータを収集した。

4.2 テストデータ

スポッティングが望まれるようなシーンでは、スポッティング対象となる連続指文字以外の手形状が多く含まれていると考えられる。そのためテストデータとして、スポッティング対象の連続疑似指文字を含みつつ、それ以外の時間は予め定めた対象とならない疑似指文字(図5)を行っている動画を撮影した。

1テストデータあたり4,000フレームの動画シーケンス(図6)を5つ撮影し、合計 $4,000 \times 5 = 20,000$ 枚の画像を用いて実験を行う。なお、実験結果の評価の簡略化のために疑似指文字の出現回数を予め定めているが、識別器に回数情報は与えられない。

4.3 実験結果

学習データから得た識別器で、テストデータに対するスポッティング性能を従来手法として単一の更新係数 α を用いた場合と提案手法として複数の α を用いた場合で比較を行った。テストデータにおけるその時点ごとの類似度の計算を行う際に、入力部分空間を2.2で挙げた更新方法を用いて、更新を行った。 α の値としては、0.1刻みで0.1から0.9までの9種類について実験を行った。評価指標として、Equal Error Rate(EER)を用いた。結果を図7に示す。

図7から全てのクラスについて、従来手法を提案手法が優

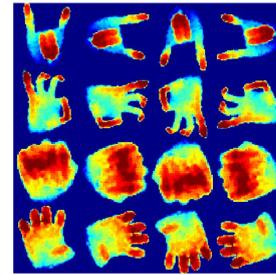


図5: テストデータに出現する辞書に含まれない疑似指文字:辞書に含まれない手形状を4つ設定し、上下左右の四方向に手を向けたものである。

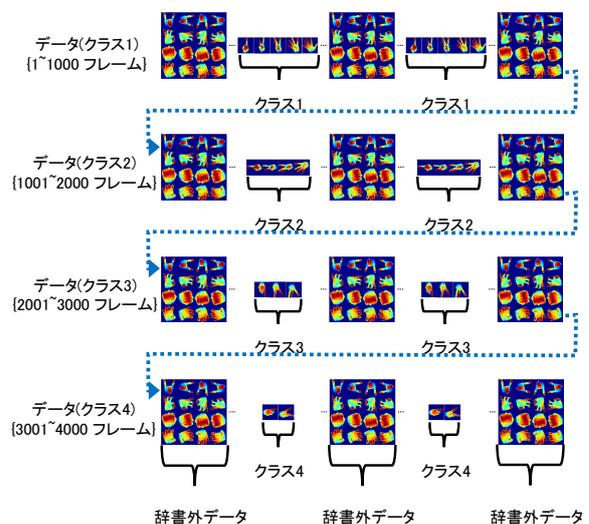


図6: テストデータの生成方法:それぞれのクラスが2度現れる1,000フレームの画像列を1試行として5回ずつ撮影。それらを全クラス1~4を昇順に試行回数が等しいものについて結合することで4,000フレームのテストデータを5試行分生成する。

れていることがわかる。加えて、従来手法の結果は最もEERの低い α を選択した結果である一方で、提案手法の結果は $\alpha = 0.1, 0.5, 0.9$ と全体を網羅するよう、適当に選択している。つまり、実験的にパラメータを最適化することなく、従来手法よりも良い結果を得られた。現実の問題では、 α を実験的に最適化することは困難であるため、望ましい性質である。

5 結論

本研究では、入力の更新を並列に行うことで、明示的な時間方向のセグメンテーションを行わずに、長さの異なる対象の識別を安定かつ正確にする手法を提案した。また、パラメータの細かな調整を必要としないという望ましい性質を持つことを確認した。

今回対象としたデータはあくまで疑似指文字であるため、こ

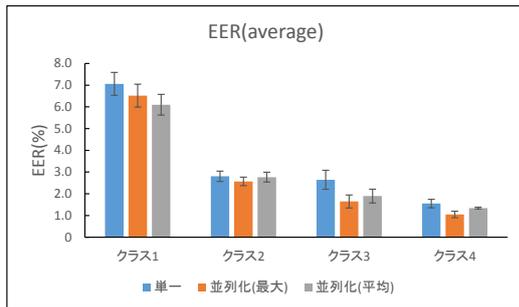


図7: スポットニング実験の EER

の実験では想定できていない問題もあると考えられる。そのため、実際の指文字を集めて実験を行うことが課題として挙げられる。実際の単語には、手の位置の移動などの情報を用いる指文字も含まれるため、手領域の画像以外の情報も組み合わせることで、より広範な指文字単語のスポットニングシステムの構築を目指す。

謝辞

本研究は科研費(25282173)の助成を受けたものである。

参考文献

- [1] 田畑 慶人, 黒田 知宏, 眞鍋 佳嗣, 千原 國宏, “手型認識を用いた指文字教育システム”, 教育システム情報学会誌, Vol.18, No.2, pp.172-178, 2001.
- [2] 平山 亮, 舟川 政博, “ニューラルネットによる静正画像からの指文字認識” 情報処理学会全国大会講演論文集 第72回平成22年(2) pp.13-14, 2010.
- [3] 高林 大輔, 大川 泰弘, 瀬戸山 浩平, 田中 陽土, 加藤 伸子, 岡崎 彰夫, 福井 和広, “フィードバック機能を備えた指文字練習システムの試作”, 信学技報, Vol. 112, No. 483, Human Information Processing, pp.79-84, 2013.
- [4] K. Fukui and O. Yamaguchi, “The Kernel Orthogonal Mutual Subspace Method and Its Application to 3D Object Recognition”, Asian Conference on Computer Vision, pp.467-476, 2007.
- [5] Y. Ohkawa and K. Fukui, “Hand Shape Recognition Using the Distributions of Multi-Viewpoint Image Sets”, IEICE Transactions, Vol.E95-D, No.6, pp.1619-1627, 2012.
- [6] O. Yamaguchi, K. Fukui, and K. Maeda, “Face recognition using temporal image sequence”, Automatic Face and Gesture Recognition, pp.318-323, 1998.
- [7] エルッキ・オヤ著, 小川 英光, 佐藤 誠 訳, “パターン認識と部分空間法”, 産業図書, 1986.