6ZB-08

音響センシングを用いた任意平面に対するジェスチャ入力

川戸　美輝 ∗　　藤波 香織 †
∗ 東京農工大学 大学院　工学府 情報工学専攻　　† 東京農工大学 大学院 工学研究院 先端情報科学部門

1　はじめに

近年，ユビキタスコンピューティング環境の普及により，コンピュータの役割は大規模なコンピューティングデバイスから日常生活のアシスタントデバイスへと変化している．したがって，コンピュータと人間を繋ぐ入出力機構において，伝統的な入出力デバイスであるキーボードやマウスを凌ぐ，より自然な相互作用をもたらすことが可能な手法が必要とされており，様々な研究が行われている．特に，三次元空間のジェスチャ認識など，筆記面を制限しない自然で新しい入力方法が注目されている．しかしながら，ウェアラブル機器やカメラを用いる手法では高い精度が期待できるものの，ウェアラブル機器の装着によるストレス [1] や「監視されている」ことに起因する不快感や煩わしさをユーザが感じることが懸念される [2]．この問題を解決するため，本稿では音響センシングを利用する事の利点に着目し，入力装置として机上面などを含む任意の平面を拡張することを目的とした入力システムを提案し，その認識精度を評価する．

2　ジェスチャ認識手法概要

2.1　処理フロー

ジェスチャ認識におけるデータの処理手順を図 1 に示す．ユーザからの入力を取得するセンサとして，音データを取得するマイクロフォンと振動データを取得する振動センサを用いる．マイクロフォンから取得した音データに対してノイズの除去を目的としたフィルタリングを行い，振動データと音データの双方で大きな変化が存在する区間をジェスチャ区間と判定して切り出しを行う．その後，特徴量を抽出し，それをもとにジェスチャを分類し，ユーザの入力を認識する．

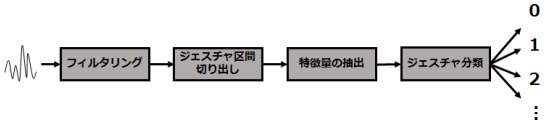

図 1: ジェスチャ認識における処理フロー

2.2　特徴量算出

時系列のデータの分類に頻繁に利用される手法として Dynamic Time Warping (DTW) があるが，分類するクラス数が増加すると十分な精度が得られない．事前の実験から，音声認識で用いられることが多い MFCC（メル周波数ケプストラム）のような周波数ベースの特徴量が有効ではないことが判明している．そのため本稿では時系列ベースで波形の情報を持つ特徴量を採用する．図 2 に例を示す．まず，波形の極大点（ピーク），極小点，中間点に注目し，それぞれの数や大きさに加え，発生した位置を全体に対する割合で表す．これにより，ジェスチャの長さによらない，大まかな波形を表現する．さらに，極大点と極小点の間の振幅を示す中間点同士の距離を計測することで，信号の立ち上がりの早さ等の波形特性が読み取れる．本稿で扱うジェスチャデータにおいては，1 画ごとに信号のまとまりが存在するため，これを分割して考えることも可能である．従って，図 2 に示すように信号の極小点を境界として前半部，後半部と分割して特徴量を算出する．

分割後のデータから抽出する特徴は，継続時間や最大，平均などの統計値を使用する．これにより，各ジェスチャの 1 画目および 2 画目以降の波形の様子が読み取れることが期待される．また，波形全体の音量の変化を表現する特徴量として，二乗平均平方根（RMS 値）を利用する．1 つのデータを 10 個のウィンドウに分割し，それぞれに対して RMS 値を算出したものを 10 次元の特徴量として採用し，大まかな音量の変化を表現する．さらに，特徴量ベースの分類器にも DTW 距離を取り入れることで，より時系列方向への柔軟性を向上させることが可能であると考える．DTW 距離の算出時の比較対象として，複数人のデータからばらつきが出ないよう，各クラス 5 つの代表点を選定した．これら代表点と比較した DTW 距離の平均を 17 次元の特徴量として採用する．

計 89 次元の特徴量の有用性の検証として，各特徴量が持つ情報量の調査を行った．その結果，データ後半部から得た統計量を始めとして，それぞれ多くの情報量を持つことが判明したため，これらの特徴量は分類モデルの構築において十分に有用であると考える．

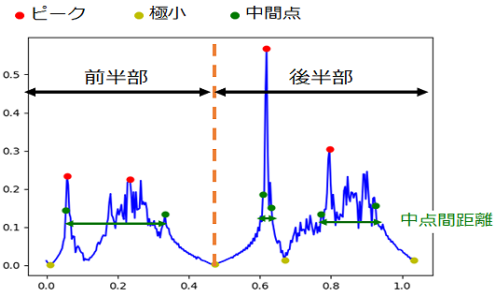

図 2: 特徴量の抽出（‘4’ の例）

2.3　分類器の構築

本稿では，分類精度の向上を目的として，分類器の階層構造化を提案する．誤分類を生じるクラス群を 1 つのグループとしてみなした状態で 1 層目の分類を行い，2 層目では誤分類を生じやすいクラス群の分類に特化した分類器を使用して詳細な分類を行うことで，モデル全体として精度の向上が見込めると考える．事前に簡易的な手法として DTW を用いて分類を行い，誤分類を生じる傾向のあるクラスを調査した．その結果判明した誤分類を生じやすいグループを以下の表 1 に示す．これらの類似したクラス群を 1 つのクラスとしてみなし，1 層目では 9 クラス分類を行う．そして，グループ 1 から 4 に分類されたサンプルは 2 層目の分類器で 2 または 6 クラスに詳細に分類することで精度の向上を試みる．

Acoustic-based Gesture Recognition on Planar Surfaces
∗ Miki KAWATO　† Kaori FUJINAMI
∗ ,† Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology

表 1: 相互に誤分類を生じるクラス群

グループ 1	0, 6	グループ 3	7, R
グループ 2	4, V	グループ 4	2, 3, 8, 9, △, 波線

3 データ収集

事前調査としてデータ収集時の条件による分類精度への影響を，簡易的な分類モデルを用いて確認した．その結果，指やペンといった入力手段に関しては分類に影響がないことが確認され，卓上に配置する物体の有無についても影響が無いことが判明した．このことから，データ収集ではよりユーザ負担の少ないペンによる入力採用し，卓上には実験機材のみを配置する．また，設置する机の材質の差異による影響は，様々な材質ごとに特化した分類モデルを用意して使い分けることで対応可能であるため，本稿においては机の天板として採用されることの多いメラミン化粧版のみを使用する．

ジェスチャ認識の分野において，アプリケーションに依存しない万能なジェスチャセットおよびデータセットは存在せず，システム設計者自らが使用する用途に応じて決定する必要がある．本稿ではユーザが好ましいと感じるジェスチャ条件について調査した研究 [3] を参考に，0 から 9 の数字，×，△，□， V，R，波線，・（ノック）から成る 17 種類のジェスチャセットを決定し，データ収集を行った．対象とする被験者は 15 名（右利き 14 名，左利き 1 名）とし，1 名につき各ジェスチャ100 サンプル，合計 25500 サンプルを収集した．

4 評価実験と考察

4.1 実験概要

本稿では，3 章で収集したデータを利用して，オフライン環境で分類精度評価を実施する．より汎用的な性能を調査するため，検証は一人抜き交差検証を用いて行う．また，階層化分類の各階層で利用する分類アルゴリズムは Random Forest（木の数は 100 本）とした．さらに，2.3 節で構築した階層型分類器だけでなく，比較対象として Random Forest (RF)，Support Vector Machine (SVM)，Convolutional Neural Network (CNN) の 3 種を用いる．RF は多クラス分類で頻繁に使用され，学習および評価が高速に行えるという利点が存在する．SVM は，伝統的な分類手法の一つであり利用される機会も多いことから比較手法として妥当であると考えられる．CNN は，深層学習の一つであり，近年注目されている手法である．従来の手法のように設計者自らが特徴量を選定する必要がなく，ある程度自動で最適な分類モデルを構築可能であるという利点がある反面，学習に大量の教師データを必要とする欠点も存在する．本稿では伝統的な特徴量を利用する手法を中心に述べてきたが，近年注目されている手法との比較として CNN を採用する．CNN は，畳み込み層 3 つとプーリング層 2 つに加え，ドロップアウト層で構成する．

4.2 実験結果

各手法を用いて検証を行った結果を以下の図 3 にまとめる．さらに，提案手法である階層化した RF による分類における Confution Matrix（混同行列）を次の図 4 に示す．4 手法それぞれの平均精度は 0.79 (0.065)，0.67 (0.081)，0.60 (0.067)，0.69 (0.083) となった（括弧内は標準偏差）．

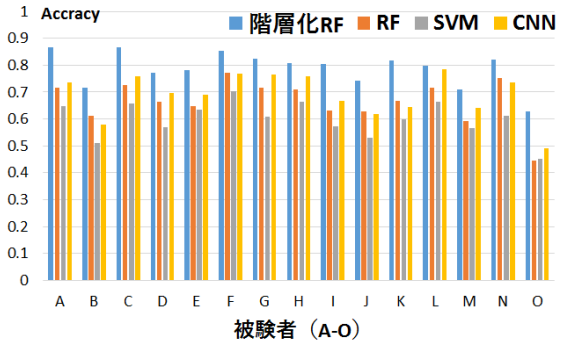

図 3: 一人抜き交差検証結果

	0	1	2	3	4	5	6	7	8	9	V	R	X	△	□	波線	・
0	709	93	190	114	37	1	57	1	66	76	49	0	0	86	0	26	0
1	56	1358	18	0	0	0	52	0	2	3	5	0	0	0	0	0	22
2	27	54	1194	13	47	10	58	0	21	23	41	0	0	21	0	6	0
3	40	2	23	1122	56	1	93	0	28	22	64	1	19	28	0	20	0
4	0	2	3	27	1172	60	0	8	2	12	2	5	192	21	8	5	0
5	0	0	26	19	53	1106	0	16	28	65	16	81	11	15	21	60	0
6	49	110	103	173	6	2	763	1	67	140	34	0	2	44	0	14	0
7	0	10	3	16	6	1	2	1392	20	5	6	14	5	28	2	5	0
8	102	2	26	27	1	2	52	7	1240	14	5	3	1	21	0	16	0
9	14	1	31	30	7	7	19	2	27	1291	33	1	9	36	0	11	2
V	101	22	138	190	1	4	91	14	69	80	766	0	8	29	0	2	0
R	1	0	1	7	0	3	0	19	11	13	0	1306	9	2	86	60	0
X	0	2	0	11	81	6	1	13	4	6	44	6	1329	9	2	2	0
△	11	6	20	17	0	1	2	56	23	17	32	0	1	1318	2	7	0
□	0	1	0	0	2	13	0	74	1	2	0	79	12	0	1312	19	0
波線	1	1	18	20	1	0	7	8	20	11	2	12	0	4	9	1401	0
・	0	10	0	0	0	0	0	0	0	0	0	0	0	3	0	1	1496

図 4: 階層化 RF の混同行列（行：正解値，列：推定値）

4.3 考察

図 3 より，全ての被験者のデータにおいて検証した 4 つの手法の中で提案手法が最も高い正解率（最高 0.87，最低 0.62）を得た．有意水準 1%のウィルコクソンの順位和検定を実施した結果，提案手法は他の 3 手法に対して有意に正解率が高いことが判明し，階層型分類器構成の有用性が確認された．本稿では，分類モデルの学習に被験者本人のデータを使用しなかったが，実際のシステムでは，ユーザ本人のデータを後から追加することで，ユーザに適した分類モデルへと改良していくことが可能であり，今回の検証結果で得られたものよりもさらに精度の向上が見込める．以上のことから，提案システムの入力システムとしての有用性も示唆される．一方，図 4 におけるクラス ‘0’ の分類精度に注目すると，他のクラスと比較して非常に低精度であることが読み取れる．このように，誤分類を生じやすいクラスの高精度での分類が行われていない．今後，第 2 層の分類において，これらのクラス群の分類に特化した特徴量を考案する必要がある．

5 おわりに

本稿では，音響センシングを用いたジェスチャ入力システムを提案し，データの収集，特徴量の算出手法などについて述べ，その認識精度について検証を行った．結果として，提案システムと分類アルゴリズムの有用性が示唆された一方で，分類モデルのさらなる改善の必要性が明らかになった．今後は，誤分類を生じるクラス群に特化した分類器の精度向上を目指す．

参考文献

[1] S.Moran, et al. Japanese salient perceptions of ubiquitous monitoring. In *Technology and Society in Asia (T&SA), 2012 IEEE Conference on*, pp. 1–5. IEEE, 2012.

[2] A.Oulasvirta, et al. Long-term effects of ubiquitous surveillance in the home. In *Proc. of UbiComp'12*, pp. 41–50. ACM, 2012.

[3] J.Kela, et al. Accelerometer-based gesture control for a design environment. *Personal and Ubiquitous Computing*, Vol. 10, No. 5, pp. 285–299, 2006.