料理中の料理動画再生制御に向けた料理工程推定手法の評価

城谷 知葵[1] Dawton, Billy[1] 石田 繁巳[2] 荒川 豊[1]

概要：新型コロナウイルスの流行による外出制限を受けて料理動画の需要が高まっているが，料理中に動画を視聴する際，動画再生端末を操作する必要がある．そこで本研究ではユーザの料理工程に合わせて料理動画をループ再生する手法を提案する．ユーザが行っている料理工程を動画内でループ再生することで動画再生端末を操作することなく料理動画を視聴することが可能となる．本稿では，この実現に向けた第一歩として料理音から料理工程を認識する手法について述べる．ユーザの料理音を取得し，既存の料理音データセットで作成した機械学習モデルに入力することで料理工程を認識する．実環境ではテレビの音やBGM が混ざる可能性があるためバックグラウンドノイズの影響について検討した結果も報告する．一般家庭で録音した料理音データを kitchen20 データセットで作成した分類モデルで初期的評価を行った結果，3 種類の料理工程を正解率 0.84 で分類できることを確認した．また料理工程が頻繁に変わらないことを考慮して平滑化処理を行った結果も示す．

1. はじめに

新型コロナウイルスの流行による外出制限を受けて自宅で料理をするための料理動画の需要が高まっている．料理動画とは具材の切り方や炒め方などの料理工程が解説されているものものであり，YouTube などの動画共有プラットフォームに数多く投稿されている．料理工程を目で見て真似できるため，料理経験が浅い人にとってそのような動画を手本にして料理することは便利である．

しかし，ユーザが動画を見ながら実際に料理をする場合には，料理工程に合わせて動画の一時停止や巻き戻しを行う必要がある．多くの料理動画はユーザの見やすさを配慮して，食材を切る・炒めるなどの各料理工程を示した短い動画によって構成されている．そのため，料理動画の進行度よりもユーザの進行度の方が遅れてしまい，一時停止や巻き戻しが必要となるが料理中は手が汚れていることもあり，スマートフォンの操作は困難を伴う．

そこで本研究ではユーザの料理行動を調理音を用いて認識し，料理動画内の対応する工程をループ再生する手法を提案する．**図 1** に提案手法の全体像を示す．例えば，ユーザが「切る」工程をしている場合，料理動画の「切る」工程をループ再生する．この時，料理動画は料理工程ごとにあらかじめセグメントされているものとする．スマートフォン端末に内蔵されたマイクでユーザの料理工程を認識し，対応する工程を料理動画内でループ再生する．

図 2 に提案手法の一連の流れを示す．本手法は 3 つのブロックで構成される．料理音取得ブロックでは，ユーザが料理動画を視聴しながら料理を行い，その音をマイクで録音する．ユーザが手軽に使用できることと，動画を視聴しながら料理をすることを想定しているため，料理音の録音はスマートフォン内蔵のマイクを用いる．分類ブロックでは，料理ブロックで得られた料理音に対して特徴量を抽出しあらかじめ作成された分類モデルに入力する．照合ブロックでは，分類ブロックで得られた料理行動と料理動画の対応する部分を照合し繰り返し再生処理を行う．料理動画内の工程は文献 [1, 2] などの既存の料理動画分析手法を用いてあらかじめラベル付けしておく．以上の一連の流れにより，ユーザが料理中に動画の一時停止や巻き戻しをすることなく料理動画を参考にすることが可能である．

本稿では，この実現に向けた第一歩としてユーザの料理音を取得する料理音取得ブロックと料理行動を認識する分類ブロックについて述べる．ユーザの料理工程を認識する手法としてカメラや加速度センサを用いた研究 [3–6] が報告されているが，ユーザにセンサを装着して料理することや，環境側に複数のセンサを設置することは一般家庭での使用を想定すると現実的ではない．料理動画をスマートフォン端末で視聴することを想定しており，スマートフォン端末内のマイクの使用は，行動認識のための新たなセン

[1] 九州大学大学院システム情報科学研究院
[2] 公立はこだて未来大学システム情報科学部

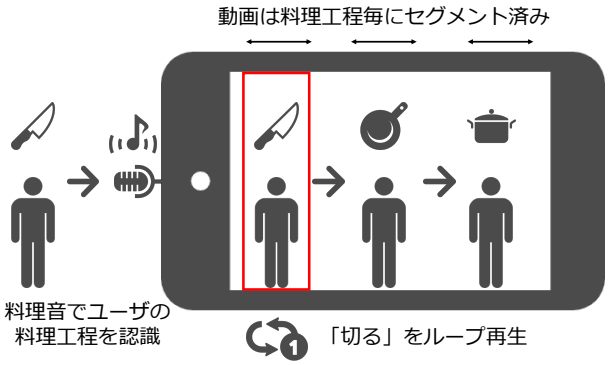


図 1　提案手法の全体像

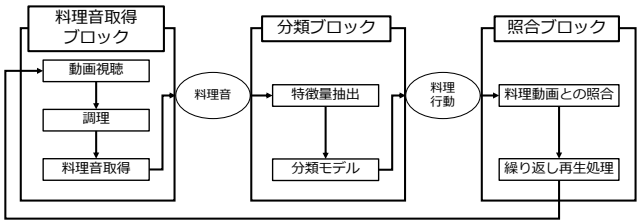


図 2　提案手法の一連の流れ

サを必要としないため非常に手軽である．現実環境では人の話し声やテレビの音，BGM などの音楽などが料理の音に混じることが予想されるため，本稿では，音を用いた料理工程の認識におけるバックグラウンドノイズの影響について検討した結果を報告する．

提案手法の実現可能性を検証するため，一般家庭のキッチンに iPhone11 を設置して肉じゃがを作った際の料理音を用いて初期的評価を行った．実環境では音楽やテレビの音が背景で流れている可能性もあるため，分割した音にノイズとして BGM を合成し，BGM が料理工程推定性能に与える影響を評価した．その結果，切る・炒める・煮るの 3 種類の料理工程を正解率 0.84 で分類できることを確認した．

本稿の構成は以下の通りである．2 章では関連研究を示し，3 章で提案する手法を示す．4 章で初期的評価を行い，最後に 5 章でまとめとする．

2. 関連研究

本章ではまずセンサで取得したデータによる料理工程認識手法について述べ，次いで音データを用いた研究について述べる．

2.1 料理工程認識手法

料理工程を認識する手法として，カメラや加速度センサを用いた研究が広く行われている．文献 [3] では，頭部にカメラを固定し「切る」・「混ぜる」など計 5 種類の料理工程を認識している．調理工程の認識率は平均正解率 0.63 であり，調理器具の情報も加味すると平均 0.84 の精度を実現している．文献 [4] では，キッチンにカメラを設置し，関節の動きを用いることで 65 種類の料理行動を平均正解率 0.58 で認識している．文献 [5] では，ユーザの手首に加速度センサを装着し「切る」動作を正解率 0.82 で認識している．

カメラと加速度センサを組み合わせた例もある．Cooking Activity Recognition Challenge [7] は料理中の行動認識に関するコンペティションであり，慣性データとモーションキャプチャデータを使用している．このコンペティションで 1 位を獲得した手法 [6] では，0.95 の正解率で料理行動を分類できている．しかし，カメラや加速度センサを用いた手法はユーザにデバイスを装着することや，大規模なデバイスを環境側に設置することが必要であるため，ユーザへの負担や導入コストが高いことが懸念される．

マイクを用いた料理行動認識手法も提案されている．文献 [8] では，スマートフォンのマイクを用いて焼きそばを調理した際の音を録音し，「切る」・「焼く」・「その他」分類を最大で F 値 0.71 の精度で実現している．カメラと加速度を組み合わせた手法に比べて精度は落ちるが，ユーザの身体にデバイスを装着する必要はなく，スマートフォンで手軽に行えるという利点がある．

しかしながら，マイクを用いた料理行動認識手法を家庭内で用いる場合，生活ノイズの影響を受ける．家庭内ではテレビの音や BGM などの料理音以外の生活音があり，その音は家庭によって異なる．複数人で生活している環境であれば話し声がノイズとして含まれる可能性もある．

本研究では，手軽にユーザの行動を認識したいためマイクを使用し，家庭内のノイズにロバストなモデルを作成することを目標とする．

2.2 特徴量

料理音の機械学習ではメル周波数ケプストラム係数（MFCC）や log-mel スペクトログラムなどの特徴量が用いられている [8–10]．

文献 [8] では MFCC を用いている．MFCC とは人間の聴覚特性を反映し，音声認識でよく用いられている特徴量である．次元数を削減できることから，料理音などの環境音に対してもよく用いられている．

近年ではニューラルネットワークを用いて環境音を分類する手法が盛んに行われている．文献 [9, 10] は log-mel スペクトログラムを使用している．log-mel スペクトログラムは MFCC の前段の特徴量であり，ニューラルネットワークと相性が良い．

本稿ではデータ数の観点からニューラルネットワークは使用しないため，MFCC を特徴量として使用する．

3. 提案手法

図 3 に提案手法の概要を示す．料理音取得ブロックでは音を取得するだけであるため，本項では主に分類ブロック

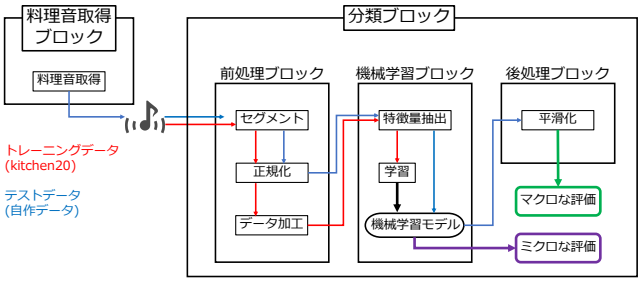
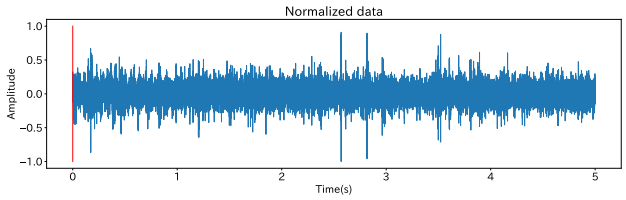
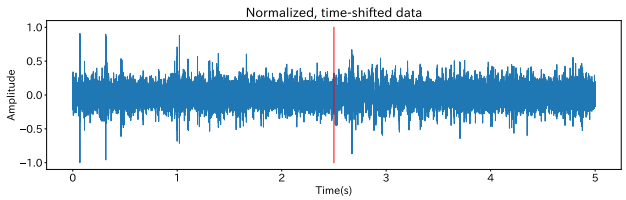


図 3　提案手法の概要図



(a) 音データの波形



(b) 2.5 秒のタイムシフトを施した音データの波形

図 4　タイムシフトを施した波形

について述べる．分類ブロックは前処理ブロック，機械学習ブロック，後処理ブロックの 3 つのサブブロックに分割できる．

3.1　前処理ブロック

前処理ブロックでは，音データに対してセグメント・正規化・データ加工の 3 つの処理を施す．

まず，音データを固定時間幅のウィンドウに分割する．音データから特徴量を抽出する際，データ長を揃える必要がある．モデル作成に用いる kitchen20 データセット [11] の各音源の長さは 5 秒であるため，ここではウィンドウサイズは 5 秒とした．

次に正規化を行う．本稿では音データの大きさを −1 から 1 までの範囲に正規化する．

最後に，ロバスト性を高めるために，タイムシフト・ホワイトノイズ付加・BGM 付加の 3 つのデータ加工を各ウィンドウに施す．

タイムシフト

図 4 にタイムシフトを施した波形を示す．赤い線が元々の音データの開始時刻である．開始時刻をずらすことで音の発生タイミングをずらした新しいデータを作成する．本稿では各ウィンドウで開始時間を半分遅らせ，ずれた後半のデータは前に加える処理をする．

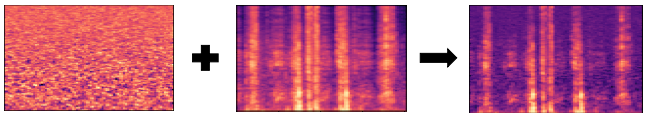


図 5　ホワイトノイズを付加したスペクトログラム

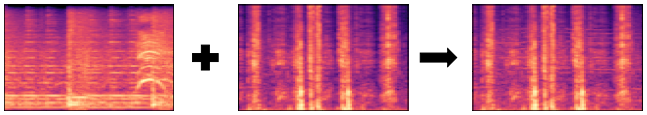


図 6　BGM を付加したスペクトログラム

ホワイトノイズ付加

図 5 にホワイトノイズとホワイトノイズを付加した音データのスペクトログラムを示す．スペクトログラムとは音データの情報を示した図のことであり，横軸が時間・縦軸が周波数・色の濃淡がパワーを示している．ホワイトノイズは正規分布に従うノイズであり，音データに付加することによって機械学習の認識精度をあげることが期待できる．

BGM 付加

図 6 に BGM と BGM を付加した音データのスペクトログラムを示す．実環境では BGM が流れている可能性があるため BGM をノイズとして付加する．本稿では GTZAN データセット[*1]の “rock” ジャンルの 19 番目のデータを用いた．GTZAN データセットの音源の長さは 30 秒であるため最初の 5 秒間のみ使用した．

3.2　機械学習ブロック

機械学習ブロックでは，前処理ブロックで処理された音データに対して特徴量を抽出し料理行動分類モデルを作成する．

まず，前処理ブロックの出力データから特徴量を抽出する．2 章でも述べたように，環境音分析によく用いられている MFCC を利用する．MFCC を導出するための短時間フーリエ変換のパラメータとして，窓関数はハニング窓，FFT サイズは 2048 サンプル，ホップ幅は 512 サンプルとする．MFCC は第 1 成分から第 20 成分の計 20 次元を用いる．短時間フーリエ変換の各ウィンドウで得られた各 MFCC について，平均値・最大値・最小値・分散を計算して特徴量とした．

次に，特徴量をもとに機械学習を行う．本稿では分類モデルは限定しないが，LightGBM を用いた．分類モデルによって出力された料理工程は後処理ブロックへの入力として用いられる．

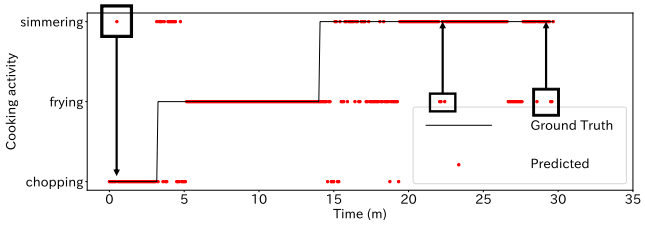


図 7　後処理の概要

3.3　後処理ブロック

図 7 に後処理の概要を示す．後処理ブロックでは，料理工程が頻繁に変動することは少ないと仮定し，機械学習ブロックで得られた料理工程に対して平滑化を施す．例えば，食材を切っている最中に数秒間だけ煮ることは現実的ではないため，分類モデルがそのような結果を出力した場合は誤分類である可能性が高い．誤分類を減らすため，t 秒間連続で同じ工程が続かない場合は，1 つ前に分類した工程と同一と見なす処理を施す．

4.　評価

提案手法のロバスト性を検証するため，公開データセット及び実験的に取得したデータを用いて評価を行った．

4.1　機械学習モデルのトレーニング

3.2 で述べた機械学習モデルの学習には公開データセットである kitchen20 データセット [11] を用いた．kitchen20 データセットから “frying-pan”・“chopping”・“boiling-water” のラベルが付いたものを取り出し，3.1 章で述べたデータ加工を施した上で学習を行った．予備実験より，各トレーニングデータセットに対してホワイトノイズの SN 比が 26dB，BGM の SN 比が 20dB になるように付加した．前処理のデータ加工は 3 種類あるため，元のデータセットに加えて 3 種類それぞれのデータ加工を行ったデータセットを作成し，その全て，すなわち元のデータの 4 倍のデータを使って機械学習モデルを学習した．

4.2　テストデータ収集実験

テストデータは実際に調理をする実験を行って，その際の音を取得した．**図 8** に実験環境を示す．一般家庭のキッチンに iPhone11 を設置して被験者 1 名が肉じゃがを作った際の料理音をサンプリング周波数 48kHz，AAC 圧縮形式で録音した．実験時には，テレビや BGM などの料理音以外の音は発生させず，静かな環境で録音した．肉じゃがを料理する際の工程は以下の通りである．

(1) じゃがいも・にんじん・玉ねぎを包丁で切る．
(2) 上記の野菜と豚肉をフライパンで焼く．

*1 GTZAN Dataset – Music Genre Classification, https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

図 8　実験環境

表 1　使用したデータセットの個数

	kitchen20	テストデータセット
切る	45 × 4 個	39 個
焼く	68 × 4 個	130 個
煮る	46 × 4 個	188 個
その他	0 個	109 個
合計	159 × 4 個	466 個

(3) 水と調味料を加えて煮る．

野菜の皮を剥いたり洗ったりする「その他」の処理は (1) の工程に含まれている．テストデータセットには工程を手動でラベル付けし，各工程ごとに長さ 5 秒にセグメントした．

表 1 に使用したデータセットの個数を示す．各データセットは 1 個当たり 5 秒間の音データである．

4.3　ミクロな評価

ミクロな評価では分類モデルが出力した料理工程に対しての混同行列で評価した．テストデータセットには実験で取得した肉じゃが音を使用した．なお，野菜の皮を剥いたり洗ったりする「その他」の工程はミクロな評価では「切る」工程から手動で除外した．

図 9 にテストデータセットの前処理を示す．実環境では音楽やテレビの音が背景で流れている可能性もあるため，BGM をノイズとして付加した．使用した BGM は GTZAN データセットの “disco”・“pop”・“metal”・“blues”・“rock” であり，各音楽ジャンルデータセットから上記の順番で 1 つずつデータを抽出し BGM データセットを作成した．BGM データセットの音量を −1 から 1 の間で正規化し，テストデータとの SN 比を −10dB から 10dB まで変更してテストデータセットに 1 つずつ付加した．最後に，BGM を付加したテストデータセットを −1 から 1 の間で正規化し，各料理工程のデータセットを「切る」データセットの個数である 39 個になるようにランダムアンダーサンプリングして評価した．

表 2 に SN 比を変更した時の正解率と各料理工程の適合率・再現率を示す．表より，料理音に対して BGM が大きくなる，すなわち SN 比が小さくなると精度（正解率・適

表 2　SN 比を変更した場合のミクロな評価結果

SN 比 (dB)	正解率	切る		焼く		煮る	
		適合率	再現率	適合率	再現率	適合率	再現率
−10	0.63	0.87	0.33	0.93	0.64	0.48	0.92
−8	0.68	0.93	0.36	0.91	0.74	0.51	0.92
−6	0.68	0.94	0.38	0.84	0.79	0.52	0.85
−4	0.71	0.89	0.41	0.89	0.82	0.56	0.90
−2	0.72	0.86	0.46	0.89	0.79	0.57	0.90
0	0.74	0.91	0.51	0.89	0.82	0.59	0.90
2	0.74	0.87	0.51	0.89	0.79	0.59	0.90
4	0.75	0.92	0.56	0.89	0.79	0.60	0.90
6	0.80	0.90	0.69	0.91	0.79	0.68	0.92
8	0.80	0.88	0.74	0.89	0.79	0.69	0.87
10	0.84	0.87	0.85	0.89	0.79	0.77	0.87

表 3　閾値 t を変更した場合の累計正解時間

	切る (秒)	焼く (秒)	煮る (秒)	その他 (秒)	合計 (秒)
真値	195	650	940	545	2330
$t=0$	165	540	745	0	1450
$t=10$	190	535	775	0	1500
$t=15$	195	535	825	0	1555
$t=20$	195	535	805	0	1535

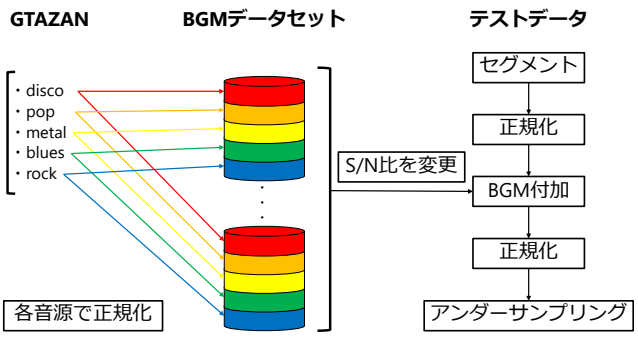


図 9　テストデータの前処理

合率・再現率）が低下する傾向があることがわかる．特に「切る」に対しての再現率は BGM の音量による影響が大きく，SN 比が −10dB の時は 0.33 となっている．しかし，SN 比が 10dB である時の全体的な精度は BGM を付加していない状態と比較して極端に低下していない．実環境では，小さい音量であれば BGM などのノイズが録音時に流れていてもその影響はないと言える．

図 10 に，料理工程分類結果の混同行列を示す．図 10(a), (b) はそれぞれ SN 比が ∞，−10dB のときの結果を示している．図 10(a) の SN 比が ∞ の場合は BGM を付加していない場合であるが，「煮る」以外の工程は高い精度で識別できたことがわかる．同じ調理器具を使用する「煮る」音は「焼く」音に誤分類するケースが比較的多かった．SN 比が低下して −10dB となると，図 10(b) に示すように「切る」の多くが「煮る」に誤分類されていることがわかる．「煮る」と BGM は定常的な音である一方で，「切る」は瞬間的に発生する音である．このため，大きな BGM を定常的な「煮る」音に誤分類してしまったと考えられる．ノイズ環境下での識別精度向上に向けては，瞬間的に発生する「切る」のような音の特徴を捉えられる特徴量を追加する必要があると考えられる．

図 11 は，機械学習モデルのトレーニング時に 3.1 に示したデータ加工を一切行わなかったモデルで料理工程を推定した結果の混同行列を示している．図は，テストデータの SN 比が 10dB の場合を示しており，正解率は 0.79 であった．表 2 の結果と比べると，正解率が 0.05 低下していることがわかる．テストデータの SN 比が 0dB，−10dB の場合の正解率はそれぞれ 0.70，0.58 であり，いずれの場合も提案手法より 0.04 程度低下していることがわかる．提案手法では 3 種類のデータ加工を行うことで実環境で付加される音に対するロバスト性を高められたと考えられる．

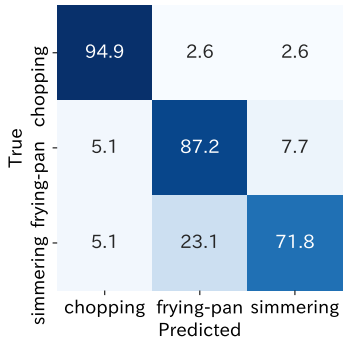
4.4　マクロな評価

マクロな評価では各料理工程の Ground Truth とモデルの分類結果を時間軸上で比較したグラフで評価する．テストデータセットはランダムアンダーサンプリングせず全て用いる．マクロな評価では「切る」工程における野菜の皮を剥いたり洗ったりする「その他」の工程も含めた．

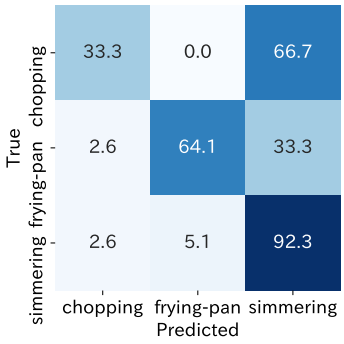
図 12(a) にマクロな評価を示す．黒い線は Ground Truth，赤点はモデルの出力を時系列に沿ってプロットしたものである．先に述べた通り Ground Truth には「その他」の項目が含まれているが，トレーニングデータには加えていないため分類モデルは「その他」の分類はできない．図 12(a) より，後処理を施していないマクロな結果では料理工程が頻繁に変動していることがわかる．

図 12(b) に $t=15$ で平滑化した場合の結果を示す．平滑化を施す前の図 12(a) と比較すると散らばっていた赤点が固まり，変動が少なくなっていることがわかる．

次に，閾値 t の平滑化性能への影響の評価として，平滑化後の分類結果が Ground Truth と一致している累計正解時間を評価した．**表 3** に，閾値 t を変更した場合の累計正解時間を示す．平滑化前（$t=0$）の結果と比較すると，「その他」以外の 3 つの料理工程において平滑化により累計正解時間が増加していることがわかる．閾値 $t=15$ の時が累計正解時間の合計がもっとも長い結果となった．閾値 t が長いとユーザの料理工程が切り替わった際の動画の切り替えに遅れが生じるというデメリットが存在する．このため，ループ再生システムの実装時に閾値に関するアンケー



(a) SN 比 $= \infty$ のとき



(b) SN 比 $= -10$dB のとき

図 10 料理工程分類結果の混同行列

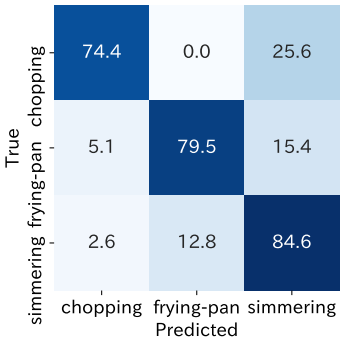
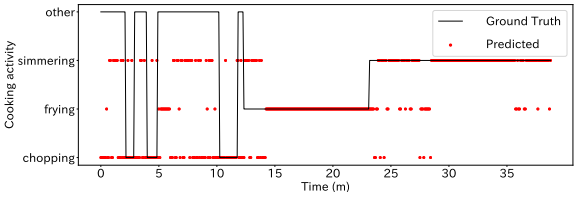
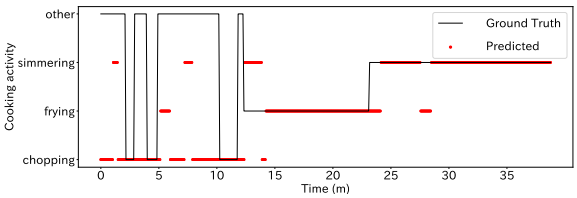


図 11 前処理をしないデータで学習した場合の混同行列



(a) 平滑化前の図



(b) 平滑化後の図（$t = 15$）

図 12 平滑化前後の図

トを取得するなどの必要があると考えられる．以上より，ループ再生システムを実装する際，ユーザが料理している間はユーザと同一の料理工程をループ再生できるが，ユーザの料理工程が移り変わった際に動画に数十秒の遅れが生じる可能性があることがわかった．

5. おわりに

本稿では，ユーザの料理行動を調理音を用いて認識し，料理動画内の対応する工程をループ再生する手法を提案した．その実現に向けた第一歩として，スマートフォン端末で録音したユーザの料理音データから料理行動を認識する手法を示した．実環境で録音した料理音データを用いて初期的評価を行い，ノイズや BGM を加えたモデルを用いることで録音環境が異なる場合でもロバスト性を高められること，工程の推定結果に対して平滑化を行うことで推定精度を向上できることを確認した．

謝辞 本稿で示した研究の一部は，科研費 (JP19KT0020, JP21K11847) 及び東北大学電気通信研究所共同プロジェクト研究の助成で行われた．

参考文献

[1] Bianco, S., Ciocca, G., Napoletano, P., Schettini, R., Margherita, R., Marini, G. and Pantaleo, G.: Cooking action recognition with iVAT: an interactive video annotation tool, *Int. Conf. Image Analysis and Processing*, Springer, pp. 631–641 (2013).

[2] Zhang, H., Lai, P.-J., Paul, S., Kothawade, S. and Nikolaidis, S.: Learning collaborative action plans from YouTube videos, *Int. Symp. Robot. Res.(ISRR)*, pp. 1–16 (2019).

[3] Ooi, S., Ikegaya, T. and Sano, M.: Cooking Behavior Recognition Using Egocentric Vision for Cooking Navigation, *Journal of Robotics and Mechatronics*, Vol. 29, No. 4, pp. 728–736 (2017).

[4] Rohrbach, M., Amin, S., Andriluka, M. and Schiele, B.: A database for fine grained activity detection of cooking activities, *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1194–1201 (2012).

[5] 広井芳秋ほか：空間メモリを利用した調理動作認識に基づく調理プロセスの観測，大学院研究年報理工学研究科編，Vol. 45 (2015).

[6] Picard, C., Janko, V., Reščič, N., Gjoreski, M. and Luštrek, M.: Identification of Cooking Preparation Using Motion Capture Data: A Submission to the Cooking Activity Recognition Challenge, *Human Activity Recognition Challenge*, Springer, pp. 103–113 (2021).

[7] Cooking Activity Recognition Challenge, `https://abc-research.github.io/cook2020/`.

[8] Korematsu, Y., Saito, D. and Minematsu, N.: Cooking State Recognition based on Acoustic Event Detection, *The Workshop on Multimedia for Cooking and Eating Activities*, pp. 41–44 (2019).

[9] Adaimi, R., Yong, H. and Thomaz, E.: Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions, *The ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5, No. 1, pp. 1–24 (2021).

[10] Liang, D. and Thomaz, E.: Audio-based activities of daily living (ADL) recognition with large-scale acoustic embeddings from online videos, *The ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3, No. 1, pp. 1–18 (2019).

[11] Moreaux, M., Ortiz, M. G., Ferrané, I. and Lerasle, F.: Benchmark for kitchen20, a daily life dataset for audio-based human action recognition, *Int. Conf. Content-Based Multimedia Indexing (CBMI)*, IEEE, pp. 1–6 (2019).